DOCUMENT RESUME

ED 050 176 TM 000 580

AUTHOR	McLaughlin, Kenneth F.
TITLE	Interpretation of Test Results.
INSTITUTION	Department of Health, Education, and Welfare, Washington, D.C. Office of the Commissioner of Education.
REPORT NO	Bull-7; OE-25038
PUB DATE	64
NOTE	66p.
EDRS PRICE	EDRS Price MF-$0.65 HC-$3.29
DESCRIPTORS	Achievement Tests, Aptitude Tests, *Counseling, Expectancy Tables, Group Tests, Guides, Individual Tests, Intelligence Tests, Item Analysis, Multiple Choice Tests, *Parent Counseling, *Standardized Tests, Test Construction, Testing Programs, *Test Interpretation, Test Reliability, *Test Results

ABSTRACT

This bulletin attempts to explain the use and limitations of regularly the process of selecting educational areas which should receive additional interpret adequately their meaning to parents and students. A companion publication "Understanding Testing Purposes and Interpretations for Pupil Development," also prepared by HEW, was issued in 1960. A general discussion of the development of a standardized test is followed by consideration of specific types of tests, including intelligence or scholastic aptitude tests and achievement tests. Scoring a multiple-choice type test, the accuracy of test results, and the analysis of class achievement are also discussed. A section on classroom interpretation of test scores provides helpful suggestions on how to handle the interpretation of this material with students and parents. An extensive list of selected references is included. (TA)

ED050176

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY

OE-25038
Bulletin 1964, No. 7

Interpretation of Test Results

by Kenneth F. McLaughlin
Specialist, Appraisal of the Individual

U.S. DEPARTMENT OF
HEALTH, EDUCATION, AND WELFARE
Anthony J. Celebrezze, *Secretary*

Office of Education, Francis Keppel, *Commissioner*

Printed 1964
Reprinted 1965

Superintendent of Documents Catalog No. FS5:2 25:25038

U.S. GOVERNMENT PRINTING OFFICE
WASHINGTON: 1965

For sale by the Superintendent of Documents, U.S. Government Printing Office
Washington, D.C. 20402 Price 30 cents

Foreword

UNDER TITLE V, Guidance, Counseling, and Testing of the *National Defense Education Act of 1958*, the Congress of the United States has recognized the value of tests as a tool which may be used to help make an early determination of the aptitudes and abilities of the students in our schools. This bulletin attempts to explain the use and limitation of regularly administered tests, so as to enable administrators, counselors and teachers to interpret better their meaning to parents and students.

Participation of a student in a testing program and the recording of the test scores on his cumulative record are not sufficient. Each counselor or teacher who works with a student, as well as the student himself, should know the student's strong and weak points—the strong points in order to develop them further, the weak ones in order to recognize limitations and to determine where extra effort must be applied. Test results, properly interpreted, can be of great assistance to all concerned with the instruction of youth.

A companion publication, *Understanding Testing, Purposes and Interpretations for Pupil Development*, was issued in 1960 (OE–25003). Both of these publications have been prepared by the Guidance and Counseling Programs Branch.

ARTHUR L. HARRIS
Associate Commissioner
Bureau of Educational
Assistance Programs

Contents

	Page
Foreword	III
I. Introduction	1
II. Development of a Standardized Test	5
Tests as Samples	5
Construction of a Standardized Test	5
Use of Test Results	7
III. Intelligence, Mental Ability, or Scholastic Aptitude Tests	9
IV. Achievement Tests	13
V. Scoring a Multiple-Choice Test	15
VI. Accuracy of Test Results	17
VII. Analysis of Class Achievement	21
Scattergram Analysis	21
Prediction Tables	27
Error Analysis:	
Made Outside of the Classroom	31
Made Inside the Classroom	34
Item Analysis Methods	34
"High-Low"	35
"Alternate Response"	38
By Test Scoring Machine	39
By Typewriter	43
VIII. Classroom Interpretation of Test Scores	47
IX. Interpretation of Test Results to Parents	53
Individual Conferences	53
Group Conferences	54
X. Selected References	59

Contents

Figures

		Page
1.	Method for Plotting Pairs of Percentiles on a Scattergram	23
2.	Method for Plotting Pairs of Percentiles on a Scattergram for a Single Class	25
3.	Expectancy Table for Predicting the Probability that a Student with a Certain Grade-Point Average in the Ninth Grade Will Obtain a Certain Grade-Point Average in the Tenth Grade	29
4.	Table Showing Errors Made on a Test of Arithmetic Concepts	32
5.	Examples of High-Low Item Analysis [N=36]	37
6.	Sample Item Analysis Sheet	40
7.	Sample of an Item Analysis by Typewriter With a Detailed Analysis of Item 15 [N=36]	45
8.	Profile Sheet for Sam Smith of Patunka High School, Grade 10	50

I. Introduction

IDENTIFICATION OF HUMAN ABILITIES or aptitudes is not easy. Tests can be helpful, but their results alone will not provide the information needed to solve all problems or to answer all questions concerning abilities or future areas of occupational success. Test scores can give suggestions as to a level of ability which may suggest areas of success, but it must be understood that even the best predictor can only be considered as indicating "a likelihood of success" or "the odds in favor of success." For example, one might say that a youth with a certain score on a specified test has three chances to one that he may *succeed* as an engineer. However, this means also that there is still one chance in four that he will *fail* in this particular area.

For many years teachers have designed a small number of tests and administered them to their students in order to evaluate day-by-day learning. Occasionally, some of these tests have been given orally to only one student individually; some have been administered to groups of students in only one or two classes. Today, however, a standardized test [1] or a battery of tests is frequently administered to all of the students in a large number of classes in several grades in a single school, or in all grades in every school within a school system, or within a political subdivision such as a county or a State. To make the best use of the minutes or hours scheduled for such standardized testing, there must be careful preplanning by the teacher and cooperation by the student.

Periodic testing periods permit the student to evaluate his accomplishments, to determine possible weak points in one or more areas, and to compare himself with the average for other students of a similar grade or age. Availability of test scores soon after the administration of a test—

---

[1] A standardized test is a measuring instrument designed for a specific purpose. It has been carefully constructed with the cooperation of master teachers, subject-matter specialists, and test technicians. It must be administered under prescribed conditions and scored in a predetermined manner. It must be interpreted in terms of the appropriate norms which have been developed for a described population of a specified age or educational level.

1. Encourages the *teacher* to examine closely the current learning level of the student so that the course of study may be adapted to individual needs.
2. Permits the *principal* to ascertain the average class level of the students in his school so that he can determine if previously established goals are being attained.
3. Helps the *superintendent* to obtain objective information which may be used as a basis for research in curriculum development.
4. Suggests to *school board members* whether the curriculum is meeting the needs of its students.
5. Provides objective data which can inform the *community* of the accomplishments of its students as compared with nationwide averages of students in similar grades throughout the country.

It is sometimes said that a teacher after having a student in class for several months can tell as much about his academic ability as can be learned from the results of tests given at the beginning of the fall term. However, it should not be necessary for a teacher to wait a number of weeks before acquiring the information necessary to continue instruction at a child's current learning level. Early objective test results are subject to immediate verification by comparison with actual classroom accomplishment.

Relatively few teachers can recognize all the able individuals in their classes. Russell and Cronbach [2] referred to a study in which a psychologist asked each of 6,000 teachers to name the "most intelligent" child in his class. It was found that, on the basis of other evidence available about the child, only 15 percent of the teachers made a correct choice. Of course, they may have recognized other students with high potential, but they nevertheless failed to identify many of the best students.

Different teachers of the same subject have different grading standards. It would be difficult to compare one student with another from class to class or from year to year without using some common measuring instrument. A well-constructed standardized test, properly administered, can reveal in a minimum amount of time a great deal about a student's aptitudes, current achievement level, or interests.

Results of tests given to all students in a grade within a school have proved useful, along with other information from the cumulative record, as a means for grouping students with similar abili-

---

[2] Russell, Roger W. *and* Cronbach, Lee J. Report of Testimony at a Congressional Hearing (to the Senate Committee on Labor and Public Welfare on Feb. 27, 1958). *The American Psychologist* 13: 219–220, March 1958.

ties. Many school administrators believe that a teacher should be able to accomplish more with his students if most of them have about the same level of ability. With students at a similar level, the teacher does not have to restrain the bright ones while drilling the slow, or lose contact with the slow while trying to interest and anticipate the sharper questions of the bright.

It is not to be inferred, however, that *all* of the same students should be kept together for *all* classes because of the results of one general aptitude test. In many situations, for example, it might be best to separate the students in classes of arithmetic, English, or science. Even though students may be roughly grouped in the assignment to a particular class section, there may still be a wide variation in ability within each class. It is at this point that a careful analysis of test results would help the teacher diagnose quickly the weak and strong points of each student.

The administration of tests is not an end in itself. Tests should never be given simply for the sake of filling in the blanks on a student's cumulative record card. Each test should be administered for a specific purpose and used to help the student determine his educational or vocational goals. The results of standardized tests can be helpful to the student, his parents, and his teachers, as together they plan a worthwhile school program.

II. Development of a Standardized Test

IT WOULD BE DIFFICULT to design a 40- to 90-minute test which would cover completely any particular field of knowledge. For example, how could a single 40-minute test include all that a student should know about English literature? (It is not disputed that in a few cases the test *would* provide a 100 percent sample of the student's knowledge.) Or how could a teacher, in 90 minutes, test for student understanding of all of the theorems of a plane geometry course?

Tests as Samples

Since complete test coverage of a subject is not possible, it becomes necessary to take a sample of all possible items in a specified course or in a particular subject-matter area. This can be done fairly well by a classroom teacher if he follows certain procedures during several succeeding semesters. However, a test publisher has already completed such procedures when he has constructed a standardized test designed to measure achievement in a specified area. Further, the test publisher has spent many months and thousands of dollars in completing the processes necessary to make available a test which meets the consumer's requirements for reliability, validity, and norms.

Construction of a Standardized Test

One of the advantages of a standardized test [1] is that a professional testmaker constructs it according to subject specifications determined by a committee of experts in a particular subject. The test agency selects these experts from the appropriate academic level—elementary, secondary, or college. This committee, after examining numerous textbooks and courses of study from many parts of the country, determines those topics common to most of the curricula for the particular subject and grade.

The committee develops a table of specifications, or predetermined "skeleton" of topics, in outline form. It decides what

[1] McLaughlin, Kenneth F. How Is a Test Built? (U.S. Office of Education.) *Understanding Testing, Purposes and Interpretations for Pupil Development.* Washington: U.S. Government Printing Office, 1962. (OE-25003) p. 4-7.

proportion of items in the total test should be assigned to each topic in order to give a reasonable balance, based upon the varying and relative importance of the different subtopics. Members of the committee and other specialists write a large number of test items in the appropriate form to fit the predetermined outline. Objective test items may appear in any one of a number of forms, such as true-false, matching, or multiple-choice. For most subjects the item writers put the items in a four-choice or five-choice multiple form.

The committee sorts all related or similar items and uses its best judgment to select the required items for each main topic or subtopic of its outline. During this process the committee may assemble several parallel test forms.

Next, as a pretest or tryout, the testmaker arranges to administer the test forms in representative schools to a sample of students of the age or grade for which the test is designed.

After scoring, the committee analyzes each test item to determine its *difficulty;* that is, the percent of students who marked each item correctly. The committee rejects any item which all students mark correctly or incorrectly since it would have no effect on the relative ranking of each student.

After placing the students' papers in order from high to low, the committee selects a high and low group of papers, and checks each item for its *discriminating power;* that is, the percent of pupils with high total scores answering the item correctly is compared with the percent of low-scoring pupils choosing the correct answer. If the item is a good one (i.e. discriminates), more students in the high group than in the low group should mark the correct response.

Next, the committee checks to discover whether or not some of the students in the sample chose each of the distractors, or incorrect choices. If no student, or a very small number, selected a distractor, a member of the committee writes a new one to use in the next tryout of the item.

The committee selects the items which meet the required standards of difficulty and discrimination and assembles the needed final test forms. The testmaker then administers these tests to a national sample of students and establishes national norms.

If a teacher completes an analysis of the items on a standardized test, he will discover that in a small class some items may not discriminate, and a few items may be too easy or too difficult. Such information is a useful indicator of the coverage of his

course as compared with other courses in the country. The real teaching purpose of such an analysis, however, is to use the test as a diagnostic instrument. The teacher discovers how many students missed an item in a particular section of a course and can reteach these concepts.

No classroom teacher has the facilities to complete all of the steps of an item analysis *before* administering a test of his own construction to one of his classes for the first time. However, he can make a table of specifications and write items to fit the purposes of the course. After the first administration of the test, he can also complete an item analysis on his test which will point out the students' errors and will help the teacher improve test items which he may use in future tests. Various methods for obtaining item analysis information will be suggested in later sections of this bulletin.

Because the curriculum for each particular subject may vary from school system to school system, it is generally recommended that, before a standardized test is selected for use in a school, a committee of teachers in the subject-matter area examine several of the available tests to determine which one most closely fits the local curriculum. If this is not done, test scores may not be as high as expected. If the tests administered are too difficult, some of the students may feel they are not progressing as they should, and those with the lowest scores may have an unwarranted feeling of failure and lack of progress. On the other hand, if the test is too easy for the group, some students may receive such high ratings that they may become overconfident.

The teacher and administrator must understand that the norms accompanying a standardized test may be based upon a population which differs from that of their school. Whether or not this is true may be determined by examining the test interpretation manual.

Use of Test Results

Teachers and parents sometimes expect a test to diagnose all difficulties or point out a well-defined road that the student can follow until he reaches his goal. However, the road is more like one found on an ocean beach. One can see where many cars have driven—but the road is a wide one. When driving along one can swing several feet to either side without difficulty and still be heading in the same general direction. Similarly, test results may indicate a desired direction, but other available information must be used to help determine the path which each student may follow

to reach his goal. A test score is *one* of the tools of guidance. It must be used in association with other information concerning the child's background, environment, strengths, and weaknesses.

III. Intelligence, Mental Ability, or Scholastic Aptitude Tests

THE PURPOSE OF intelligence, mental ability, or scholastic aptitude tests is to provide an estimate of the ability of an individual to learn or to acquire understanding. It is sometimes said that an individual who is high in such abilities is capable, among other things, of successfully coping with novel situations to which he may be subjected. Because it is rather difficult to design tests which will indicate the level of ability necessary to reason in new situations, such abilities must be measured indirectly by tests which emphasize knowledge of vocabulary, skill in the discovery of underlying patterns, and the ability to manipulate both mathematical or abstract symbols.

A group intelligence test, when administered properly, results in a raw score which must be converted to a mental age (MA) or to some other meaningful score for comparative purposes. The mental age corresponding to each score is determined by first giving the test to large samples of students of the same chronological age. Then the average score for each age is computed and a table constructed so that the teacher can determine the mental age, in years and months, corresponding to each test score. Note, however, that this averaging method for determining the MA scale immediately suggests that the "true" MA of a particular student might be a little higher or a little lower than that indicated by the table. That is, the teacher should not imply that in a group of students of the same chronological age, a student with a computed mental age of 110 months actually has a higher mental age than a student with a mental age of 108 or a lower mental age than one with a mental age of 112. If another test were given, the mental age order of these two students might be reversed. In other words, the values obtained should be used like the 1/4-inch marks of a carpenter's rule and not like the 1/100-inch rulings on the micrometer of the machinist.

To compute the most commonly known IQ, or Intelligence Quotient, one forms a ratio, or quotient, of the mental age (MA) divided by the chronological age (CA)—this quotient being multi-

plied by 100 in order to eliminate decimal points. The preceding statement may be written as follows:

$$IQ=\frac{MA}{CA}\times 100$$

Thus, if it is determined that a student has a mental age of 10 years, or 120 months, and his chronological age is also 120 months, then the ratio of 120 divided by 120 is equal to 1. When this 1 is multiplied by 100, one obtains the ratio IQ of 100. Thus:

$$IQ=\frac{120}{120}\times 100=100$$

Again, if a child happens to have a mental age of 132 months and his chronological age is 120 months, then his IQ will be greater than 100. In this case, it would be equal to 110. That is—

$$IQ=\frac{132}{120}\times 100=110$$

Further, a child with a mental age of 96 months and a chronological age of 120 months would have a below average IQ of 80. That is—

$$IQ=\frac{96}{120}\times 100=80$$

The just described ratio IQ has several disadvantages which have been highlighted by recent research. The ratio IQ is based upon the idea that a child's rate of mental development is fixed. This has been found to be untrue. Technical characteristics of a scale related to the difficulties of the items used cause different variabilities to occur at different ages. Finally, it has been suggested that one should not apply the ratio IQ to persons over age 13.[1]

The familiar individually administered Stanford-Binet IQ was computed by the above ratio method and had a standard deviation, or variability, of 16. (This means that if the average intelligence of the whole population is considered to be 100, then the IQ's of the middle two thirds of the population would lie within a range of values from 16 points *below* 100, i.e. 84, to 16 points *above* 100, i.e. 116.) The revised (1960) edition of this test and some of the more recent intelligence tests have reported results in terms of "deviation IQ's." Under this method the mean score for a particular age has been considered to be an IQ of

[1] Cronbach, Lee J. *Essentials of Psychological Testing.* 2d ed. New York, Harper & Bros., 1960. p. 171.

100, and whatever MA falls at a position of one standard deviation above the mean for each age may be converted to an IQ of 116, if there is a desire to establish a correspondence with the Stanford-Binet. If all intelligence test scores were converted in terms of deviation scores with a standard deviation of 16, there would be less difficulty interpreting the many IQ's now appearing in transfer students' cumulative records based upon different IQ tests.

However, all of the intelligence tests developed by different publishers have *not* been equated in terms of the above standard score scale. Further differences in the meaning of IQ scores occur because the norms are based upon different samples of the population and give different mental ages. It is possible for a pupil to have an IQ of, for example, 120 according to one test and an IQ of 112 according to another. Another pupil might have an IQ of 92 on the first of these same two tests, and an IQ of 100 on the other. Thus, the counselor who uses these results, or interprets them to teachers and parents, must always know the name of each test used. He can then make his own mental correction or adjustment so that the results become more meaningful. The counselor should also know when each of the several IQ tests was given, so that he can note discrepancies or expected differences which may have occurred. Therefore, the complete name of each test and its date of administration should always be entered in each student's cumulative record. It would also be helpful to know whether the test was administered by a teacher, principal, psychometrist, or a school psychologist. Then, if there seems to be any discrepancies between the test scores, the interpreter might immediately recognize the source of unusual error.

Because of the misunderstandings which have arisen over the meaning and use of the IQ, many schools are currently administering scholastic aptitude tests rather than IQ or intelligence tests. Results cannot be reported in terms of an IQ. The report of a scholastic aptitude test is most often in terms of a percentile rank. The percentile rank is the percent of scores in a national or local distribution of scores which is equal to or lower than the score corresponding to the given rank. Thus, if a student's percentile rank on a test is 75, then his score is equal to or better than 75 percent of those scores made on the same test in either the national or local distribution.

Most of the current scholastic aptitude tests include at least two kinds of items—verbal and quantitative. Sometimes some of the quantitative items might be considered as verbal items because so-

called arithmetic "story" problems are often included. Naturally, the student must be able to read the problem in order to analyze it and arrive at a solution. It is entirely possible that a student who has a poor verbal facility and a high mathematical facility might receive a lower score than he deserves. However, most tests of this type will give a verbal and quantitative score, as well as a score based upon a composite of the two parts, so that the area of strength or weakness may be determined or further explored.

Sometimes the statement is made that mental ability tests given in the lower grades are not valid because the children are too young. However, it must be remembered that, in establishing the norms for these grade levels, other children of the same ages took the test under similar conditions. Thus, the results serve to give a general idea of the capabilities of a student.

In most school systems where a planned testing program has been established, it is customary for groups to take scholastic aptitude or intelligence tests at regular intervals of 2 or 3 years. In some schools, the same test or a higher level of the same test series is used. In other schools, it is the policy to use a different mental ability test at predetermined intervals.

IV. Achievement Tests

SCHOLASTIC APTITUDE TESTS often serve as predictors of future achievement, while achievement tests measure the actual skills or subject-matter content acquired at any grade level. At the elementary level, achievement tests measure the attainments in the basic skills areas, such as reading, arithmetic computation, map interpretation, and spelling. At the secondary level, achievement tests measure attainment in such areas as English, social science, natural science, mathematics, and foreign languages.

By incorporating carefully graded materials, a number of the available achievement test batteries cover a wide range of grade levels, beginning at grade 3 or 4 and continuing through high school or the freshmen year of college. Since it is difficult to cover satisfactorily such a large grade range with a single test, a series of tests has been developed in each subject, each test covering several grades, such as grades 4–6, grades 6–8, and so on. When the grades tested are overlapped with two tests, the teacher has several choices. For example, if a teacher has an advanced 6th grade, he might give a test covering grades 6–8; if the group is slow, he may choose the test covering grades 4–6. Other achievement batteries cover either the elementary or secondary school grades—but not both. If a school uses parallel forms of the same battery at frequent intervals—that is, annually or biennially—it is possible to observe the growth of the student in each of the areas included in the battery.

Test results from a coordinated testing program are most important to the teacher as he tries to group his students, or to discover the weaknesses of each student or of each class in the various subject-matter areas. Summary record charts are often available from the test publisher for recording certain combinations of scholastic aptitude and achievement tests. These forms may be designed to show a student's academic growth profile or to show class strengths or weaknesses. Similar charts can be made by the school or by the teacher to fit the chosen tests. A study of these charts by a test specialist or counselor may suggest irregularities which have occurred in the administration of the test. For example, if all of the scores of an average class

seem to be much higher or lower than would be expected, one might consider whether too much or too little time was permitted for the tests, whether the teacher gave extra help to a class, or whether a teacher failed to follow an important instruction for a particular class.

Since achievement tests can be helpful to teachers, it is important that such tests be selected with care. Before a final choice is made, the test content should be compared with the appropriate curriculum to determine whether the items included are covered in the local program and would be fair to the students.

V. Scoring a Multiple-Choice Test

THERE ARE a number of ways to score a multiple-choice test, whether it be a standardized test or a teacher-made test. One of the earliest and most widely known methods for scoring a specifically designed 8½"×11" answer sheet which has been marked with an electrographic (current-conducting) lead pencil is by means of the IBM 805 test-scoring machine. A punched answer key, or matrix, which is inserted in the machine permits a small unit current to flow for each correct answer marked by the student. The bits of current are added to give a reading on a meter dial which indicates the total number of correct answers. If a special scoring formula is required, the machine, when properly set, will automatically deduct a fraction of a point for each incorrect answer.

New electronic scoring machines which are located in several test-scoring centers require that the student make an opaque mark in the required space on a different type of answer sheet. Then an optical scanner, which "reads" these "spots," automatically records the total number of correct answers. The answers to as many as nine tests of a battery can be marked on the two sides of a special answer sheet, along with the student's coded name. The machine will "read" the name and will score these papers at the rate of more than 6,000 papers an hour. A computer is used to determine the percentile or standard score corresponding to each raw score.

Some test companies have developed answer cards which can be scored by special mark-sensing machines or optical scanners. Such machines have the test data available immediately for further statistical analysis.

A number of new scoring machines continue to appear on the market. Since they are designed primarily to score teacher-made rather than standardized tests, these machines are small and in some cases portable. In most cases, these machines will do only one thing—give a total raw score. Thus, it is not possible for the teacher to complete an item analysis with them which will permit the use of the test results for diagnostic purposes.

One portable test-scoring machine weighing less than 25 pounds

uses a "porta-punch" type card hand-punched by the student. The operator is required to note visually the "number right" indicated by a counter mounted on the front of the machine and to write this number on the answer card before clearing the machine to score the next card.

Another type of scoring machine is the size of a duplicating machine and weighs 50 pounds. It operates automatically to score up to 200 new-type answer sheets for one loading of the machine. Special lead pencils are not required to mark the answer sheets. The number of wrong answers and omitted questions is printed automatically on the answer sheet and the questions which are missed are automatically marked on each paper. The number of questions missed by an entire class is recorded on a counter.

Although there are a number of machine procedures for test scoring, one should not neglect several of the simplest procedures which can be used when necessary with both standardized and teacher-made tests—hand scoring. There are several kinds of hand-scoring answer keys which may be used—such as the fan (or accordion) key, strip key, and cut-out key. Each of these keys is designed so that, if properly adjusted, the correct response for each question will appear near the designated answer space on the student's paper. The teacher can then make an accurate comparison of the answer key and the student's responses.

It is possible to punch out a blank scoring card, or matrix, to fit an answer sheet, whether it is homemade or purchased. If it is a standardized test, it is often possible to take the punched key which is provided for machine-scoring purposes and use it for hand scoring by placing it over the answer sheet and counting the correct answers. (NOTE: Some tests which have many subparts may require one set of keys for machine scoring and a different set of keys for hand scoring.) Counting correct marks by 2's, that is, 2, 4, 6, 8, is quicker than counting each correct response singly. For large scoring jobs an inclined scoring frame to hold the key and answer sheets will speed the process.

VI. Accuracy of Test Results

COUNSELORS AND TEACHERS who interpret test scores must remember that a test score does not represent a precise point on a scale. One must think of the wide mark made by a stub pencil, or an even larger interval or band, as representing the region which one is certain includes a student's "true" test score. By a "true" test score one means a number which would represent exactly the level of ability or achievement which a test is supposed to measure. It is impossible to ever find this "true" score. However, one can be reasonably certain, with known probability, that an obtained score does not differ from the "true" score by more than a certain amount.

The uncontrolled or chance "error" which is inherent in test scores is referred to as the "standard error of measurement." This means that if it were possible to administer the same test to a student several times, without any learning occurring in between, his test scores would vary by several points. Therefore, one becomes somewhat concerned as to how well a particular test score is an estimate of a student's true score. This information should be included in tables in the publisher's manual which accompanies each test. Some publishers show, for the same test, a different standard error for various parts of the score distribution.

Consider an illustration of the interpretation of the standard error of measurement as it relates to the bell-shaped distribution called the "normal" curve. The key to understanding the meaning of the standard error of measurement is to note that one determines the probability that the obtained score of a student does not miss its true value by more than a certain specific amount. In ever larger intervals, one determines this probability by multiplying the standard error of measurement by ± 1, ± 2, $\pm 5/2$ and applying values derived from a normal probability table. For example, suppose that a student's true score on a test is 75 and the standard error of measurement is 4. According to the normal probability table, the chances in this case are approximately 2 out of 3 that the obtained score does not miss its true value by more than ± 4 points. The obtained score of the student would

be somewhere in the range of 71 to 79; i.e., $75-1\times4=71$ to $75+1\times4=79$. The chances are approximately 19 out of 20 (or approximately 95 out of 100) that the obtained score lies within the range of 67 to 83; i.e., $75-2\times4=67$ to $75+2\times4=83$. Finally, the chances are approximately 99 out of 100 that the obtained score lies within the range of 65 to 85; i.e., $75-5/2\times4=65$ to $75+5/2\times4=85$. In most cases, however, it is sufficient to consider only the range of scores between plus and minus one standard error of measurement. In the usual situation, the raw score of the student is the best estimate of the true score. Thus, if the student's raw score is 75, as in the above example, we would say that the chances are 2 out of 3 that the true score would lie between 71 and 79, and so on.

In considering the standard error of measurement in terms of percentile ranks, one would generally have a numerically larger interval, or band, than that indicated by the standard error in terms of raw-score units. The percentile band will have the greatest width at the center of the score distribution, where there is the largest number of cases, and will be narrower toward the ends of the distribution. Continuing with the preceding example, suppose that a *raw* score of 75 corresponds to the 60th *percentile*, then the *percentile band* for one standard error above and below the score would be approximately from 48 to 71. For a *raw* score of 85 which corresponds to a *percentile* of 84 the *percentile band* would be from 76 to 90.

The magnitude of the standard error of measurement must be computed for each test. In some tests it may be five or six raw-score points. In others, it may be only a point or two. Its value depends upon the reliability of the test, which is determined and presented by most publishers, and the variability or standard deviation of the test scores. If, for the same class, two test-score distributions were equally variable, the standard error of measurement would be smaller for the test which is more reliable.

By using the standard error of measurement, the teacher or counselor examining the test results may know the range, or band, of possible ability or achievement suggested by each score on the test. For this reason, each teacher should read the test interpretation section of the manual which accompanies each test. Most test publishers have taken great care to compute and communicate the standard error of measurement for each test or subtest. Other information which will make test results more meaningful, such as prediction table, intercorrelation matrices, and sample applications, is also often included in the manual.

Additional errors in scores may be introduced in administering the test if the administrator does not read the manual and follow its crucial instructions. For example, the teacher must adhere *exactly* to prescribed time limits and must continually proctor the students during the testing period. When a test specialist or counselor notices that most of the test scores of a class appear to be much higher or lower than expected, he should check immediately with the test administrator to determine any irregularities in test administration which could affect the students' scores.

Another source of error is inaccurate scoring. Trained clerks are generally more accurate than teachers. For tests which must be scored by hand, every test paper should be independently scored twice, preferably by a different person each time. When the tests are scored by the IBM test-scoring machine, it is recommended that at least every tenth paper be checked a second time. Some scoring services perform the operation twice on different machines. Scoring by means of the new high-speed electronic machines is fantastically accurate.

Additional errors may occur during the conversion of a raw score into a more meaningful score, such as a percentile or a grade-equivalent. Such computations *must* be double-checked. Hand-transcription errors to cumulative or other records must be eliminated, or at least diminished, by double-checking all entries. Although high-speed electronic scoring procedures may reduce errors of scoring and norming to a minimum, scores entered by hand in the cumulative record must be checked unless individual score reports are available on the recently developed pressure-sensitive press-on labels as part of the scoring machine high-speed printer output.

In summary, before teachers, counselors, or administrators begin the interpretation of recorded test scores, they must have confidence that, except for known error, no additional errors have been introduced into the test results because of improper administration, inaccurate scoring, failure to read the appropriate norm tables, or the incorrect transcription of scores to permanent cumulative record folders.

VII. Analysis of Class Achievement

A TEACHER OR COUNSELOR knows that standardized test scores are only a portion of the many systematically recorded bits of information and observations concerning the ability and promise of a student. These data are available in the cumulative record folder of each student.

Individual test scores may be interpreted in terms of national norms, or local norms which may be based upon a class, a single school, or a complete school system. For maximum effectiveness, both standardized and teacher-made tests should be analyzed as soon as the scores are available.

By the analysis of standardized and teacher-made test results, a number of questions similar to those given below can be answered:

Is the student working up to the level of his ability?

What is the probability of student success for different subjects at the next grade level?

What kinds of items are missed most frequently on standardized tests?

What are the most common misconceptions or most frequent student errors in each class in the main divisions of each curriculum?

How can the teacher determine his best test items and the form and content of those items which need to be changed in order to give a better evaluation of each student?

There are several procedures—scattergram analysis, prediction tables, error analysis, and item analysis—which can help the counselor or teacher to answer such questions. Each of these suggested procedures can be completed in a reasonable amount of time and may be applied to either standardized or teacher-made tests.

Scattergram Analysis

Two-way charts, or scattergrams, for a class are constructed to picture for each student his relative score position (a raw score, scaled score, or percentile rank or grade) with respect to any two of the following items:

1. A scholastic aptitude test.
2. An achievement test.
3. A term or semester grade for a course in which he is currently enrolled.
4. The class grade or test score of the student at a later time in his school career.

A scholastic aptitude test often includes subscores of verbal ability and quantitative ability as well as a total score. One of these scores is often plotted on a chart along with another score or grade made by a student. The pairs of test scores for a number of students can be plotted on the same chart. For example, one could plot the verbal portion of an aptitude test together with the test score on an English or social science test or with a grade received in either of these subjects.

Similarly, the quantitative score of the aptitude test can be plotted together with a test score or grades in arithmetic, algebra, geometry, or one of the sciences. Each teacher would ordinarily plot only those scattergrams related to his own teaching field or to a subject with which many students in his homeroom class may be having difficulty. The counselor, on the other hand, can use the appropriate charts from the different fields when he talks with teachers, students, or parents.

The method for constructing a scattergram can be illustrated briefly. On large-squared graph paper, draw a vertical line near the left side of the paper superimposed on one of the printed rulings; draw a horizontal line near the bottom of the sheet joining the vertical line. These two lines, intersecting at right angles at the lower left-hand corner, are called the axes.

The scores for one type of test (e.g., an aptitude test) can be plotted with respect to one axis while scores for a second type of test with which the first is being compared (e.g., an achievement test) can be plotted with respect to the other. While it does not matter with respect to which axis either of the two sets of scores is plotted, one type of score should be consistently placed along either the horizontal or vertical axis only. If the decision is to plot scholastic aptitude scores along the horizontal axis—whether verbal, quantitative, or total—then subject-matter achievement test scores would be represented along the vertical axis.

As an example, assume that percentiles are available for each student on the verbal portion of a scholastic aptitude test and an

[1] The same general procedure would be followed if the data were available in raw scores, scaled scores, standard scores, or class grades.

achievement test in English.[1] Beginning at the point of intersection of the two axes, mark off the decile (tens) points along each axis. Percentiles on the horizontal axis then proceed from the lowest on the left to the highest on the right, while percentiles on the vertical scale will rise from the lowest to the highest.[2] Heavy rulings mark the 50th percentile on each axis.

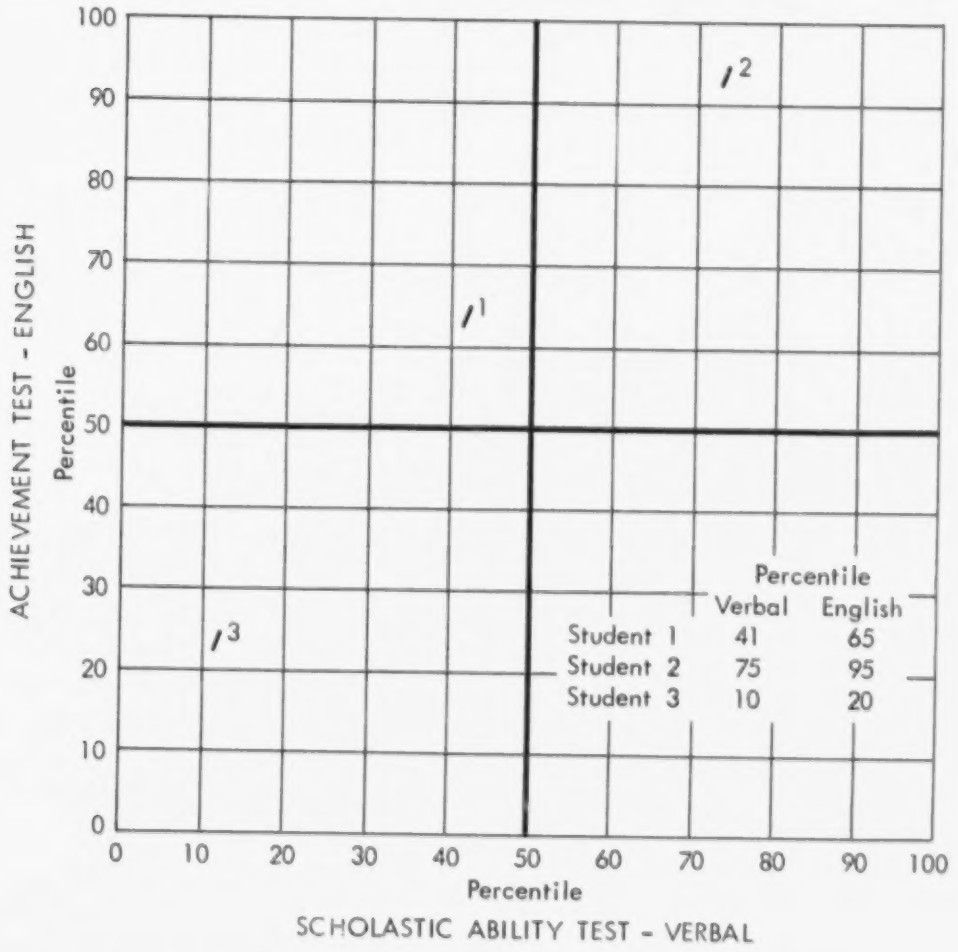


Figure 1. Method for Plotting Pairs of Percentiles on a Scattergram

There are several ways to plot the pairs of scores. One method of tabulating the numerous pairs of scores of the students in a large school, or several classes together, is to make a tally mark (/) in the appropriate square for each pair of scores. Three pairs of percentiles are plotted in figure 1. Since student 1 has a

[2] If grades are to be recorded instead of percentiles, the lowest grade, an F or a number grade, should be placed at the intersection of the axes or zero position. Letter grades on the horizontal axis must be placed in the order F D C B A so that the interpretations may apply which are suggested for scattergrams presented in terms of percentiles.

percentile of 41 on the verbal aptitude test and 65 on the English test, a tally (/) is made in the square which lies at the intersection of the vertical column between 40 and 50 and the horizontal column between 60 and 70. The plotted percentiles of student 2 are 75 on the verbal test and 95 on the English test. The plotted percentiles of student 3 are 10 and 20. (NOTE: The arbitrary rules are applied by which one plots a percentile which falls on a *vertical* line in the column to the *right* and a score which falls on a *horizontal* line in the row *above* the line, i.e., the scores 0–9, 10–19, etc., are plotted in the same column.) Such a scattergram presents a good picture of the relationships between the two tests for a large group of students.

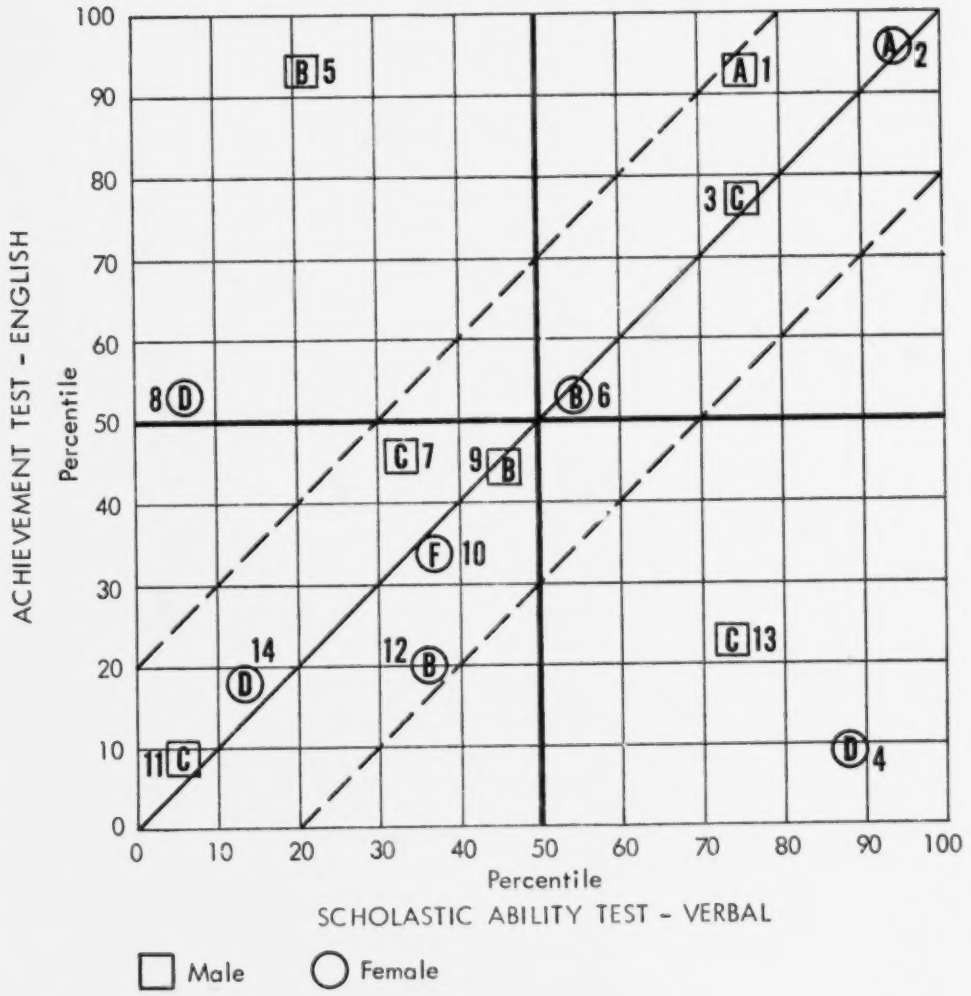
For a smaller group, such as a single classroom, another method may be used (figure 2). If all of the students in a class are listed alphabetically and assigned a number, then it is possible to identify each plotted position on the chart by placing the appropriate number beside it. In some classes indicating the sex of the student in each position may be of interest. This can be done in any one of several ways:

1. Make a square for a male and a circle for a female.
2. Assign odd numbers to the males and even numbers to the females.
3. Record the numbers in color code—blues for males and red for females.

Method 1 (used in figure 2) permits the addition of another piece of useful information to the scattergram, namely, the teacher's grade at the last marking period or at the end of the semester. As shown in figure 2, one can have the following visible information on the chart for each position: a number coded to the student, the sex of the student, a percentile on each of two tests, and a teacher's grade.

A diagonal line has been drawn in figure 2 from the lower left-hand corner to the upper right-hand corner indicating points where a student's verbal scholastic aptitude and English achievement percentiles are approximately the same. That is, a student whose plotted scores lie on this line would have such percentile scores as 25 on the verbal and 25 on the achievement test in English or 72 on the verbal and 72 on English, and so on. If each pair of scores were the same for each student, a statistician would say that there is a perfect positive relationship or a correlation of 1.00. This rarely occurs in practice.

In figure 2 a band is formed by drawing a dotted line on each side of the diagonal and parallel to it. Although the exact loca-



Note: Letter inside symbol is grade for course.
Number outside symbol identifies student.

Figure 2. Method for Plotting Pairs of Percentiles on a Scattergram for a Single Class

tion of these lines would vary somewhat with the tests or other measures used, the band plotted here suggests that consideration must be given to the fact that all scores may be subject to uncontrolled errors.

For example, a student who scores in the lower quarter of the class in scholastic ability may also score in the lower quarter of the class on the achievement test. Such a student is female No. 14 in figure 2. Since these two scores lie within the band and are of approximately the same magnitude, her achievement would be interpreted as consistent with her ability. In fact, any student

whose scores fall within the band can be considered as working at the expected level of achievement. Similarly, a student who ranks high on one test would be expected to rank high on the other, for example, female No. 2 in figure 2.

Students who are below the error band, such as No. 4 and No. 13, are probably underachievers. When students are underachieving, they should be referred for counseling and some attempt made to find the causes of the difficulty. Sometimes the student may have been ill and missed certain fundamental lessons. Since succeeding assignments assume knowledge of these basic materials, the student fails to accomplish current requirements and falls further and further behind—as shown by an achievement test. Of course, the teacher should try immediately to discover and to correct such difficulties.

Note that No. 4, who is high in verbal ability and low on the English achievement test, received a "D" at the last marking period. She is in the top sixth of her class in verbal scholastic ability but is only operating in the lower part of the bottom quarter of the class on a related English skill. Perhaps this student needs special help to improve this skill. On the other hand, it may be that she has not applied her apparently high verbal ability to school tasks and merely needs more encouragement and challenge to get down to work. Or, perhaps, some other difficulty can be discovered in an interview with her. The scattergram does *not* identify the specific difficulty a student may have, but it often calls attention to students who have problems.

Students in the upper left-hand section of figure 2 (No. 5 and No. 8) above the error band are often considered to be "overachievers," if it is possible to accept the the idea of a child "overachieving." As Froehlich and Hoyt have suggested, "It must be recognized that the term 'overachiever' is a relative concept, for no one exceeds his capabilities in achievement. As used in this discussion, it connotes that a student is achieving relatively better than others in the group with like capacity." [3]

Overachievement may occur when a student devotes an unusual amount of time and effort to schoolwork outside of school to meet certain classroom standards. Sometimes the question arises whether or not this should be encouraged if the child's health is being affected. Perhaps the child is trying to compensate for lack of acceptance among his peers by excelling in his lessons.

[3] Froehlich, Clifford P. and Hoyt, Kenneth B. *Guidance Testing and Other Student Appraisal Procedures for Teachers and Counselors.* 3d ed. Chicago, Science Research Associates, 1959. p. 163-164.

There may be a personality problem which needs attention. Or the youngster may be highly motivated to achieve because of an overwhelming interest in a given area. Again, it must be emphasized that the scattergram will *not* solve a problem, but will call attention to problem areas.

In some class situations it may be helpful to know whether there is a difference in the aptitudes and accomplishments of boys and girls. Color coding or the use of visible symbols could make such information readily available. In figure 2, a square is used to indicate a male student and a circle a female student. One might discover that the boys are doing better in science than the girls, with one or two exceptions; or the girls may excel the boys in art, but one boy may be better than the best girl in the art class.

The inclusion of the grades with the test scores in figure 2 can point up some special problems. Why did No. 3, a boy, receive a C, when he seems to have a high aptitude for English and does well on the achievement test? Does he have personality characteristics which clash with the teacher? Or his peers? Does he cause trouble in class? Does his grade include a number of nonacademic components other than measures of English accomplishment?

On the other hand, why did student No. 12, a girl, who is below the median, or 50th percentile, in scholastic ability and in the lowest quarter in English achievement, receive a grade of B? Did she really do unusually well during the last marking period because of long study hours, or were other, nonacademic, factors at work here?

It is not to be inferred from these questions that students should be graded on the basis of test scores alone. Certainly, daily work, class participation, and the quality of class projects must be included in the term mark. However, the teacher should be aware of such grading discrepancies when they occur.

Prediction Tables

The scattergram method is particularly useful as the basis for developing a prediction table for the probable success of students in a particular subject or for their probable total grade-point average in succeeding grades. As more cases are accumulated which may be used in deriving a prediction table, the better will be the prediction.

When one begins to develop prediction tables, pairs of values are needed. These pairs may consist of two test scores, two

grades, or a test score and a grade. Often such pairs of values are available for only a limited number of students. In such a situation, the first derived results must be used with caution and with an awareness that there is always a certain amount of error associated with tables of this type. However, it is helpful to have some information from which one can gain insight.

Ordinarily, in any one school the students of a particular grade are very similar to those who have passed through the school in the preceding 2 or 3 years or who will be enrolled in the following years. This assumption is fundamental in the construction and the use of prediction tables. Thus, it is important that new tables be constructed each year to include the most recent class upon which information is available. As the results of later tables are based upon larger numbers of students, the percent of probable success will tend to stabilize. This will be obvious because the prediction percents in each square, or cell, of the table may shift only a few points or not at all. If the character of the student population in a particular school changes because of econc nic or other reasons, new tables must be *immediately* computed based upon this new group.

It is impossible to compute probability tables to be used to guide those students in the *present* ninth grade as they *prepare* for tenth-grade work, unless comparable information is available on the class *currently* enrolled in the tenth grade, which is the *same* class which was enrolled in the ninth grade *last* year. If one were considering grade-point averages, this class must have already *completed* the tenth grade and the final grades must have been entered on the cumulative record cards.

As an example, suppose that two grade-point averages are available for each of 50 students in the same school for the ninth grade and the tenth grade. The counselor desires to know the probability of a student in the ninth grade with a certain grade-point average achieving success in his tenth-grade work. Since the identification of which student made a certain average is not of interest, the first step is to make tallies indicating the pairs of averages for each student in the appropriate block or cell. These tallies may be replaced by a single summary number in the proper cell, as shown in part "A" of figure 3. The ninth-grade average appears in the center of this table. The totals in the left-hand column labeled "row sum" and in the lowest row labeled "column sum" indicate the total number of cases in each row or column. The number "50" in the lower

left-hand outside corner is the sum of the rows or columns and provides a check on the number of entries.

"A" 10th grade NUMBER with each grade average							"B" 10th grade PERCENT with each grade average					
Row sum	F	D	C	B	A	9th grade average	F	D	C	B	A	Total row percent
5				2	3	A				40	60	100
10			3	6	1	B			30	60	10	100
20	1	4	10	5		C	5	20	50	25		100
10	2	3	5			D	20	30	50			100
5	2	2	1			F	40	40	20			100
50[1]	5	9	19	13	4	Column sum						

[1] Sum of row or column sums.

Grade	Grade-point average
F	= 0.00 - 0.50
D	= 0.51 - 1.50
C	= 1.51 - 2.50
B	= 2.51 - 3.50
A	= 3.51 - 4.00

Figure 3. Expectancy Table for Predicting the Probability that a Student with a Certain Grade-Point Average in the Ninth Grade Will Obtain a Certain Grade-Point Average in the Tenth Grade

The appropriate row sum is used as the divisor to determine the percents placed in corresponding cells in part "B" of figure 3. For example, the 2 under B in the top row of "A" is divided by the 5 and multiplied by 100 to give 40 percent. That is, $2/5 \times 100 = 40\%$. 40 is entered in the first row under B in part "B" of figure 3. Then the 3 under A in the first row of "A" is divided by 5 and multiplied by 100 to give 60 percent. That is, $3/5 \times 100 = 60\%$. The 60 is entered under A in the first row in part "B". The sum of 40 plus 60 gives 100, as indicated in the "total row percent" column. This number indicates that all percents for this row are probably correct. In the second row from the top of "A", the row sum is 10. Similar computations may be made as before. Thus, $3/10 \times 100 = 30\%$ which is entered under C in the second row of part "B". And so on. The right-hand column

totals of part "B" all add to 100 percent for each row, which serves as a check.[4]

Part "B" of figure 3 is read as follows: If a student made an A in the ninth grade, the chances of making an A in the tenth grade are 60 out of 100, the chances of making a B are 40 out of 100, and the chances of making a B or better are 60+40, or 100 chances in 100. If a student made a C in the ninth grade, the chances of earning an F in the tenth grade are 5 in 100; of making a D are 20 in 100; of making a C are 50 in 100; of making a B are 25 in 100 (or 1 chance in 4); of making a C or better is 50+25 or 75 in 100 (or 3 chances out of 4); etc.

A table similar to that just described could be constructed by any counselor on the basis of high school grade averages of all college-going seniors from his school who have completed 1 college year and for whom first-year college-grade averages are available. (The high school grade average would be represented by the middle column between "A" and "B".) One must recognize the inherent inaccuracy of such a table when one combines college freshmen grade averages from many different schools with varying standards. Part of this error can be eliminated, however, by constructing a table based upon those students who have entered and remained in a single nearby college.

It is possible to construct any number of related scattergrams such as English grades in high school versus English grades in college, or English test scores in the tenth grade in high school versus the course grades in college freshman English, or the grades in high school mathematics versus the grades made in the college freshman mathematics course. A study of such relationships might suggest curriculum changes to the school staff or course changes for students in a college-preparatory curriculum.

Similar tables based upon scattergrams could be produced by the registrar or admissions officer of a university or college in order to predict a college grade-point average on the basis of high school grades or scores on required admissions tests. With such a procedure, it would be possible to accumulate data on a number of different high schools and to predict the probability of success for students from each high school. Such tables could be constructed annually so that the information for each school

[4] In actual practice, one should round each cell entry to the nearest percent. In this case, the total row percent may become 99 or 101, unless arbitrary adjustments are made so that the total is 100. In making such adjustments a change in the largest percent value would give one lesser relative error. For example, adjust from 56 to 55 rather than from a 4 to a 3.

would be current and any changes or trends could be noted. A table based upon the data accumulated for several years could be developed and should be more stable than the data based upon a single year. It would be a service for the high school counselor if the college would circulate such tables to each high school for which a table is constructed. With the automatic data processing equipment now available in many institutions of higher education, such tables as the foregoing would be relatively easy to develop. Some of the college admission testing programs are now making available such information to the high schools and to the institutions of higher education.

Some high schools and many universities give orientation period tests to all entering students. These test results can be related to freshman grade-point averages or to specific course grades. Such derived predictions of high school or college success should be made available to the appropriate counselors.

A collection of probability or prediction tables of the types suggested here would be most helpful to the high school counselor. He could use them to help a student become aware of his probability of success at a particular college. It is possible that a student who would be at almost the bottom of his entering class at one of the highly selective private universities could well be at the middle or much above average in his class at another institution. The student should then be able to make an "educated guess" as to the school where he has a good chance of being admitted and where he would be challenged to do his best work.

Error Analysis Made Outside of the Classroom

Most standardized achievement tests are designed to cover a rather broad area of a subject-matter field. A diagnostic test, on the other hand, is constructed so that certain important sections of the subject are covered a number of times from different points of view in an attempt to define specific areas of deficiency. Instruction in such areas can then be emphasized further by the classroom teacher. It is often possible, however, to use an achievement test for diagnostic purposes. The procedure described here can also be used with a teacher-made test.

The principal function of an error analysis is to obtain a summary picture of the items missed most frequently by the class. At the same time, it is possible to note which students miss or omit certain types of items.

In setting up the table of errors, a teacher should examine each test item in order to determine the topic being covered. The time

needed to determine the topic for each test item can be shortened by having several teachers work together. For some standardized tests it is possible to use the analysis of items which the publisher may have included with the administration manual, the answer key, or the interpretation manual. At least one publisher includes a short topical description with a carbon-marked duplicate answer sheet designed for the teacher's and student's use. At least one test-scoring firm includes an error analysis as one of its service options. By means of currently available electronic data-processing systems, it would be possible to group the analysis of similar test items in adjoining columns on the report sheet.

If a teacher must design his own table of errors, he may find it helpful to identify groups of related item numbers by using a

Topic or Concept

Question	1 Counting	2 Relative sizes	3 Averaging	4 Reducing fractions	5 Averaging	6 Rounding		34 Estimating quantities	35 Angles	36 Averaging	Error total	Omit total
Adams, A. A.	✓			0				✓			11	3
Anderfeld, J.	0		✓	✓		✓		✓		✓	25	4
Blackster, M.	✓			✓	0			0	✓	0	15	6
Clarksen, W.			0	0				0	✓			
Smith, J.	✓					✓				0	10	3
Tinker, T.			✓	✓	0				✓	0	15	8
Williamson, K.	✓		0	✓				✓		0	18	6
Errors	4		2	7	15	3		21	25	4	310	
Omits	1		4	2	3			10	3	8		81

✓ Errors 0 Omits ▭ Related items

Figure 4. Table Showing Errors Made on a Test of Arithmetic Concepts

color code or light shading. In figure 4 which represents part of a table of errors, one group of related items has been shaded.

The "Error total" column on the right side of the table indicates how many questions were marked incorrectly by the student while the "Omit total" column indicates how many items were not attempted. Since the teacher is aware of the amount of time allotted for a test and whether or not each student had ample time to complete the test, he can judge whether omitted items are an indication of a lack of knowledge or a shortage of time. Since the total number of correct items is recorded on the student's answer sheet, it is not necessary to indicate the total number of correct items. This total could be easily obtained by subtracting the error total plus the omit total from the number of questions in the test. Since the totals given in figure 4 are for a complete table, the reader may not be able to verify all totals.

The "Errors" row at the bottom of the table indicates the number of *students missing each question*. For example, 15 students of this class of 36 students missed item 5 on averaging and 3 students omitted it. A relatively few students missed the other items on averaging questions 3 and 36. However, 8 students omitted question 36. As a teacher one would be concerned because of the large number of errors in question 5 and the number of omits for question 36.

Questions 34 and 35 are of interest here since more than half of the class tried the items and missed them, while a number of others decided to omit them—especially item 34. There are several possible explanations. First, the topics were difficult for the students. Second, these topics had not yet been presented, but were to be included later in the course. This latter explanation may be especially true if a standardized test is administered early in the fall or in the middle of the school year, or if this test is constructed to cover the work of several years.

With an analysis such as that suggested by figure 4, the teacher can quickly determine the areas of strength for the class and the most commonly missed items. Students with common areas of weakness can be given special instruction in small groups. There should be a minimum amount of class time spent discussing questions missed by only a few students. Questions on previously presented concepts which were missed or omitted by a large number of students must be examined again. However, items which anticipate topics to be developed in a later part of the course should be considered at the appropriate time.

Error Analysis Made Inside the Classroom

Paul B. Diederich suggests that an error analysis of a test can be done during classroom time by having each pupil *watch* a paper other than his own.[5] If the teacher is only interested in an overall "error analysis," i.e., in how many pupils chose any one of the *wrong* responses to a test question, then the teacher only needs to call out, "Item 1, 'b' is the correct answer. Each of you holding a paper in which item 1 was *missed*, raise your hand." Then he, or a class monitor, can quickly count the raised hands, record the number beside the test question, and proceed to the other questions. Thus, in a few minutes, the items missed by the greatest number have been identified. After the papers are returned to the students, the teacher can quickly go over those questions which were missed most often and explain why they are incorrect.

Item Analysis Methods

There are several methods for analyzing objective test results which make it possible to determine one or more of the following points:

1. *Difficulty* of an item—The percent of the students of the class answering the question correctly.
2. *Discriminating power* of the correct answer—The capacity of an item to distinguish between good and poor students; the percent of the highest scoring students answering the question correctly as compared with the percent of the lowest ranking students answering the question correctly.
3. *Effectiveness* of each response for each test item—The *number* of students selecting each response (each response should be chosen at least once).
4. *Identification* of each student making a correct or incorrect choice for each item—Permits an individually designed corrective procedure for each student.

In a few school systems it is now possible to carry out an item analysis entirely by means of an attachment to a testscoring machine or by the use of automatic data processing equipment. In other schools where such services are not available it may be necessary to use other methods. In fact, much student interest may be aroused by carrying out such procedures during the classroom period when the scored papers are returned. It has been found that pupils at all grade levels, from the primary grades

[5] Diederich, Paul B. *Short-cut Statistics for Teacher-made Tests.* Princeton, N.J.: Educational Testing Service. Evaluation and Advisory Service Series, No. 5, 1960, p. 3.

through graduate school, cooperate willingly. The students are interested in learning how many of their peers missed each item, why they made an incorrect choice, and the best answer for each question. If such an analysis has been completed for the teacher's own objective test, he immediately has information which can assist him to improve his test items for future use. He can then build up a test file of items of a known quality and difficulty which will discriminate between his good and poor students.

"High-Low" Analysis.—For some tests the teacher will find it helpful to use the classroom procedure which Diederich calls a "high-low" type of item analysis.[6] This method will reveal both the difficulty and the discriminating power of each item.

To determine the discriminating power of an item, it is necessary to split the class into two sections—those with high scores and those with low scores. The separation point is the middle or median score for the class. To find the median score the following steps are necessary. Determine the range of scores of the class, that is, the highest and lowest scores, and record them at the top and bottom of the blackboard. Write all possible scores occurring in this interval in a column, beginning with the highest score at the top of the board and continuing to the lowest. Divide the number of class members by two to determine how many papers must be tallied in order to find the middle one. Beginning with the highest score, ask how many students made each score and record the results. As soon as the cumulative total number of papers equals half the class, the middle score can be determined without completing the distribution of scores.

If there are several students' papers at this middle score, collect these papers first. Then collect all papers in two groups—those above the middle score and those below. Distribute all papers above the median score on one side of the room, and those below the median score on the other side. Then assign the several papers with the median score to the high and low side at random so that the total number of papers on each side is the same. If there should be an odd number of papers in the class so that they cannot be evenly divided, the discarding of the one paper remaining will leave one student to act as a recording monitor at the board.

It is possible to get a certain amount of teamwork in this operation if a captain is appointed for each of the two groups. The teacher, or the class member with no paper, can write the

[6] Ibid., p. 3-10.

question numbers in a column on the board and make 4 column headings:

H L H+L H–L

These headings stand for:

H—the number of the "high" group who mark the item correctly

L—the number of the "low" group who mark the item correctly

H+L—"difficulty index," the total number who marked the item correctly

H–L—"discrimination index," how many more of the "high" group than of the "low" group marked the item correctly

When the teacher asks, "How many have item No. 1 correct?" each student with the correct answer on the paper he is watching raises his hand. The captain of the high group calls his number—the "H" score. The captain of the low group calls his number—the "L" score. These two numbers are written on the board and then the recorder computes and calls out the two scores for "H+L" and "H–L".

These four numbers are always obtained in the same order. Each student writes these four numbers on the answer sheet below each question as it is computed by the board monitor. Each member of the class checks on the sum and difference. With a little practice, Diederich [7] says, this item analysis can be carried out for a one-period test in about 10 to 20 minutes depending on the number of items. This is much faster than the operation could be completed by the teacher. At the same time, an excellent learning situation develops since each student becomes involved in the test results for the class as a whole and wishes to know why he has missed some of the items.

If an item is acceptable for inclusion in later tests "high-low" differences should be equal to at least 10 percent of the size of the class.[8] For example, with a class of 36 the differences should be equal to at least 4. However, because of the large value of the "standard error," an item the "true" difference of which would turn out to be 6, might in some cases give a value of less than 4. In other words, if the difference is small, one should examine the item closely. If it seems to be a well-constructed item, it should be retained. Diederich suggests that "not more than a fifth of the items in the final test should fall below the suggested standard

[7] Ibid., p. 7.
[8] Ibid., p. 8.

and the *average* high-low difference should be above 10 percent of the class, preferably 15 percent or more." [9]

The H+L number, which indicates the total number of students choosing the correct answer, indicates the *difficulty* of the item for the class. The larger the number, the easier the item. In most cases, an item which 90 percent of the class marks correctly is too easy. On the other hand, if less than 30 percent of the class marks it correctly, it is probably too difficult.[10]

Occasionally, especially with a teacher-made objective test, a greater number in the low group will obtain the correct answer than in the high group. Then the H–L becomes negative, as in question 4 in figure 5, which is called "negative discrimination." When this occurs, the item needs further investigation. Careful examination of such an item may reveal that a few changes will improve it so that it need not be discarded. To determine what changes are necessary, the teacher might ask each member of the class why he chose one of the incorrect responses, and determine whether or not the key response was poorly written. For example, the correct response of the answer key might not attract the better students if some of the supposed incorrect choices, or distractors, were actually correct. A rewritten item may be placed in the teacher's item file and tried again in a later examination.

Question	H	L	H+L	H–L
1	18	18	36	0
2	16	4	20	12
3	13	9	22	4
4	7	13	20	–6
5	9	7	16	2
6	5	7	12	–2
7	9	9	18	0
8	6	2	8	4

Figure 5. Examples of High-Low Item Analysis [N=36]

In figure 5 the results of the analysis of several test questions are given for a class with 36 students. Item 1 is an easy item (H+L=36), since all members of the high and low group marked it correctly. It has the highest possible difficulty index—36—which indicates an easy item. (The lower the H+L score, the

[9] Ibid., p. 9.
[10] Ibid., p. 8.

more difficult the item.) Since all students in each half marked the correct answer, it certainly will have no influence in discriminating between the high and low groups. Unless one desires to begin the test with an easy item, this item would not be used in another test.

Item 7 is harder than item 1, with a difficulty index of 18. Since H=9 and L=9, H–L is 0. Therefore, this item will not discriminate between the two groups and would not be used in its present form.

Item 2 is of average difficulty and is the most discriminating item illustrated, with H–L=12.

Items 3 and 8 just barely meet the criteria for the level of discrimination (H–L) with the suggested value of 4 (i.e., 10 percent of 36 is 3.6, which is rounded to 4). Item 8 is more difficult than item 3, as shown by the indices of 8 and 22, respectively. In fact, a test should not include many items as difficult as item 8. The teacher might examine this item to determine whether it is measuring a fundamental concept which must be taught again, or if it is referring to an insignificant detail which should not have been included.

Items 4 and 6 are examples of "negative discrimination." More students in the *lower* group selected the right answer than in the upper group. Although the difficulty indices suggest that the items are not easy, these items should be rejected until they are examined and rewritten.

Item 5 is more difficult than questions 1 through 4, however, since the discrimination index is only 2, it would not be used in future tests without some revision.

"Alternate Response" Analysis.—The alternate responses, or choices, prepared for multiple-choice items often include those responses which students have been known to make most often in short-answer or free-response questions. For example, in mathematics or science the most frequent incorrect answer choices are those which would result if common errors were made in arriving at a solution. (In order to prevent a student from spending too much time on a problem, the last choice is often "none of the above.") The teacher may be more interested in the *kinds* of student errors than he is in knowing merely that a certain number of students missed a question. In this situation, the analysis would be carried out in this manner by the teacher: "Question No. 1—How many students selected choice 1?" (pause and record), "How many students selected choice 2?" (pause and record), and so on, for each of the choices for each question.

Since in most cases the majority of the class will choose the correct response, the response count takes only a few minutes.

Item Analysis by Test Scoring Machine.—If a school system or a school has the IBM 805 Test Scoring Machine, it may have available the attachment called the Graphic Item Counter. This attachment provides one of the quickest and most accurate ways for making an item analysis. After separating the scored test papers into upper and lower groups on the basis of the total test scores, the machine operator can obtain the number of students in each group marking each response to each question. This information can be obtained for 18 5-choice questions at one time, since there are 90 counters available. If 4-choice, 3-choice, or 2-choice questions are asked, one run of the answer sheets through the machine will handle 22, 30, or 45 questions, respectively. If one wishes to learn only how many students answered each question correctly, as many as 90 questions may be analyzed at one time.

The procedures suggested before are for use with a single class or a department in one school. In developing and standardizing a new test, more cases would be needed than those of a single classroom and the procedures should be followed which are described briefly in *Understanding Testing* [11] or given in detail in *Educational Measurement.*[12] In making an item analysis for a single classroom, it seems appropriate to divide the class into halves—upper half and lower half. If an item analysis is based upon a test administration to 400 or more students, then the upper and lower 27 percent of the total group will give the best results.

An Item Analysis Sheet can be mimeographed with the headings and form given in figure 6. By using legal size paper, it is possible to analyze 10 questions in each column.

The figures for the "No." columns under "Upper Group" and "Lower Group" are obtained directly from the Graphic Item Count Record. The "No." under "TOTAL GROUP" is the sum of the quantities under "No." in the Upper Group and Lower Group. The percents are obtained by dividing the recorded numbers by the number in the upper or lower groups and in the total group. An example will make these calculations clear.

Suppose that there are 40 students in a class and the division into halves places 20 students in the Upper Group and 20 students in the Lower Group. In item 1, choice 1 was marked by 15 stu-

[11] McLaughlin, Kenneth F. How Is a Test Built? In *Understanding Testing.* Washington: U.S. Government Printing Office, 1962, p. 4–7. U.S. Office of Education, OE–25003.
[12] Davis, Frederick B. Item Selection Techniques. In *Educational Measurement.* Washington, D.C.: American Council on Education, 1951, p. 266–328.

ITEM ANALYSIS SHEET

Item No.	Choice	Upper group		Lower group		TOTAL GROUP	
		No.	%	No.	%	No.	%
1	①	15	75	5	25	20	50
	2	1	5	1	5	2	5
	3	3	15	7	35	10	25
	4	1	5	5	25	6	15
	5			2	10	2	5
2	1						
	②	20	100	18	90	38	95
	3			1	5	1	3
	4						
	5			1	5	1	3
11	1	3	15	1	5	4	10
	2	1	5	2	10	3	8
	3	6	30	4	20	10	25
	4	2	10	1	5	3	8
✓✓	⑤	8	40	12	60	20	50
12✓	1						
	2			1	5	1	3
	③	8	40	6	30	14	35
X	4	10	50	12	60	22	55
	5	1	5			1	3

◯ Correct item choice

✓ Item discrimination less than desired

✓✓ "Negative" discrimination

X A "large number" choose the same incorrect choice

Example: Choice 1 is the correct answer for item No.1. 75% of the upper group and 25% of the lower group choose choice 1. These values lie in the 20-80 range, requiring a difference of 15% or more to be acceptable (75-25=50). Therefore, the item discriminates satisfactorily. The total group % for the correct answer, choice 1, is 50. Hence the difficulty index is 50%.

Requirements for Satisfactory Item Discrimination (for correct choice)

Range of values (Upper group and lower group)	Difference (Upper group minus lower group)
90-100	5 or more
80-90	10 or more
20-80	15 or more
10-20	10 or more
0-10	5 or more

Figure 6. Sample Item Analysis Sheet

dents in the Upper Group and 5 students in the Lower Group. In the TOTAL GROUP, 20 (15 plus 5) marked choice 1. Choice 2 was marked by 1 student in the Upper Group and 1 student in the Lower Group which gives a sum of 2 for the TOTAL GROUP. This procedure continues for each choice for each item in the test.

For rapid computation one can easily construct a table of percents corresponding to the number of students in half of the total group, going from 1 (which is 5%) to 20 (which is 100%). Then one fills in the % columns in the item analysis sheet for the Upper Group and Lower Group. (If this is done with a colored pencil, later analysis will be easier. In figure 6 the % columns have been shaded.) In Item 1 this becomes for choice 1, 75 and 25; for choice 2, 5 and 5; for choice 3, 15 and 35; etc. The sum of the percents in either of these columns should not exceed 100 by more than 3%, which is the maximum which might occur in some classes because of rounding errors. The total may be less than 100 if one or more students omit a question.

Another table of percents should be constructed corresponding to the number of students for the total group, in this case going from 1 (which is 2.5%, rounded to 3%) to 40 (which is 100%). Then one fills in the % column under TOTAL GROUP. (One can save these tables and develop new ones as they are needed when class size changes—because of absences at test time or changes of class size in a new school year.)

It has been shown in the literature that the test best able to put a class of students in rank order is one which has item difficulties spread over most of the range, but which has an average item difficulty of 50%, with the greatest number clustering about 50%.

As the next step, examine each test item in figure 6 and code it as suggested: A circle (O) around the correct answer choice; no further mark if the item appears satisfactory; a single check (√) if the item discrimination is less than desired; a double check (√√) if there is negative discrimination; and an "X" if a "large number" select the same incorrect choice.

In item 1 the correct answer is choice 1; 75% of the Upper group and 25% of the Lower group marked it correctly. Since there is a difference of 50% (75 minus 25), which is much greater than the suggested minimum difference of 15%, this item discriminates satisfactorily and would be a good one to include in future tests—if its other responses are satisfactory. Each of the other choices was operating since each was chosen at least once by some member of the class.

Item 2, with choice 2 as the correct one is an easy item—95% of the total group of students marked it correctly. The *larger* the percent the *easier* the item. The item does discriminate satisfactorily at this level, since there is a difference of 10% (100 minus 90). Choices 1 and 4 should be reexamined, since no one

chose them. Some test constructors believe that a few easy items of this difficulty level at the beginning of a test helps to put the examinees at ease. Almost every student's score is raised one point by such an item and his relative rank may not be changed at all when one considers the complete test.

Item 11, the number of the item at the top of the second column in the Item Analysis Sheet, shows that each choice was selected by some of the students. The double check (√√) indicates that there is a "negative discrimination" with this item, which means that more students in the Lower Group chose the keyed answer than in the Upper Group. As a result, one obtains a discrimination index of minus 20% (40 minus 60). This item does not assist in ranking the students in the proper order, but rather makes the rankings less dependable. The difficulty index, as shown in the TOTAL GROUP % column is 50%, the same as item 1—but this item 11 should *not* be used. One should examine items of this type to be sure that one has not made an error in developing the answer key. Because of rounding the TOTAL GROUP % for all choices is 101.

Item 12 is an example of a question which does not discriminate at the desired level of 15% but only 10% (40 minus 30). However, if reconsideration of the item shows that it is a good item and important to the course, retain it. Choice 1 should be changed—it was so poor that no one selected it. Choices 2 and 5 are chosen by only one student each and are much weaker than choice 4. Choice 4 must be considered, since it has been marked with an "X." Why did so many students in both the Upper and Lower Groups select it? Is it the statement of a commonly accepted fallacy? Is it so ambiguous that in one sense it may really be correct? Does this question cover a basic part of the course which needs reteaching? Has this question been keyed properly? If choice 4 should be determined to be correct rather than choice 3, then one would have "negative discrimination" as in question 11. Since one student in each group omitted the question the Total Group % is 96.

Comments should be made concerning test items omitted by the student. As one becomes experienced in examining an Item Analysis Sheet, he quickly becomes aware of the few items which many students failed to answer because of the low numbers in the TOTAL GROUP % column. If the test is timed, these items would come, in most cases, near the end of the test. If they occur randomly throughout the test, the teacher should examine the

lesson plans to be certain that they have been previously covered—and then reteach them if necessary.

When the item analysis has been completed, a summary table of marks may be made of the number of single or double checks or X's. As one becomes more skillful in constructing one's own tests and using again items which have been tried out and found successful, he will discover the number of marks diminishing. However, it will be a rare occasion when, for any given class, there will be no marks. This would also be true of standardized tests which can be analyzed in a similar manner in order to discover the weak points and errors in thinking of the students.

If each item used on a test is typed or pasted on a separate card, cataloged as to topic, and the aforementioned kinds of information concerning discrimination and difficulty recorded, it is possible to build a pool of *good* items which can be used in later classes. By recording when the item is used, the repetition of the same items in succeeding terms or years can be avoided. If the foregoing analysis shows that an item is poor, it should not be used unless it is rewritten.

Item Analysis by Typewriter.—If a teacher wishes to make an item analysis himself, he can speed up the procedure by using what is called the "typewriter method." This method will be described in detail.

After sorting the papers in order, according to their scores, highest to the lowest, divide the papers into two halves at the median, as described previously for the "High-Low" Analysis. Select the pile of answer sheets for the "high" group first, still arranged with the highest score on top. Sit at a typewriter and select any set of five keys—if five-choice multiple-choice items have been used. For example, one might choose to use the keys on the typewriter with the letters or symbols at the "home position" for the right hand corresponding as follows:

Answer choice	1	2	3	4	5
Typewriter key	j	k	l	;	¢

If a student omits an item, then strike the space bar.

Beginning with the paper of the first student with the highest score under the left hand use a finger to guide down the answer column question by question. In typing with the right hand one will feel uncertain for the first two or three papers but will soon establish a typing pattern. For example, the teacher looks at the response to question 1, observes that choice 2 was selected, and types "k"; for question 2 he observes that the student selected

the last, or fifth, choice, so he types "¢"; for question 3, with choice 1 indicated type "j"; question 4 was omitted, so one uses the "space bar"; for question 5, with choice 3, type "l"; for question 6, with choice 4, type ";"; for question 7, with choice 1, type "j" as in question 3, etc. This procedure should be continued until a symbol or space has been made for one response for each item for a single student on the same line. One may also type the student's name, if desired. The responses for 30 items, including the seven above, might look like this with an extra space following question 15 being included as a tallying aid:

k¢j l;jlklj;lkj ;k¢j;kljk¢;jklj

The next highest test paper of the "high" group should be recorded in the same manner on the second line (do not double space). This routine should be continued until the responses to each question for each paper in the "high" group have been recorded. Thus, the responses of every student to each question are always in the *same* vertical column, one below the other. At the end of the high group, triple space and proceed in the same manner for the "low" group.

Experience has shown that it is sometimes helpful to space systematically for each paper as one records letters for the answers. For example, if regular IBM answer sheets are used, space after items which are multiples of 15, i.e., after items 15, 30, 45, 60, etc. If answer sheets designed for a specific standardized test are used, the spacing will vary. If one uses an answer sheet of his own construction, an appropriate place to space might be at the end of each column of answers. This provides a visual check for the end of each group of questions. If the teacher's own answer sheet has been keyed with the correct responses, double space and type it in the same relative position below the answer rows for the "high" and "low" groups.

Figure 7 shows part of an Item Analysis by Typewriter for 30 questions, with details for question 15. Note the separation of the high and low groups and that "H" and "L" are in the same order, from top to bottom, as the groups at the top of figure 7. A straightedge placed on the paper vertically and to the left or right of each question's responses permits a rapid count of the number of responses for each group which are the same as that of the answer key. The interpretations made previously for the "High-Low" Analysis now apply.

At the bottom of figure 7 it is shown that with this typewritten method it is also possible to make the "alternate response" analy-

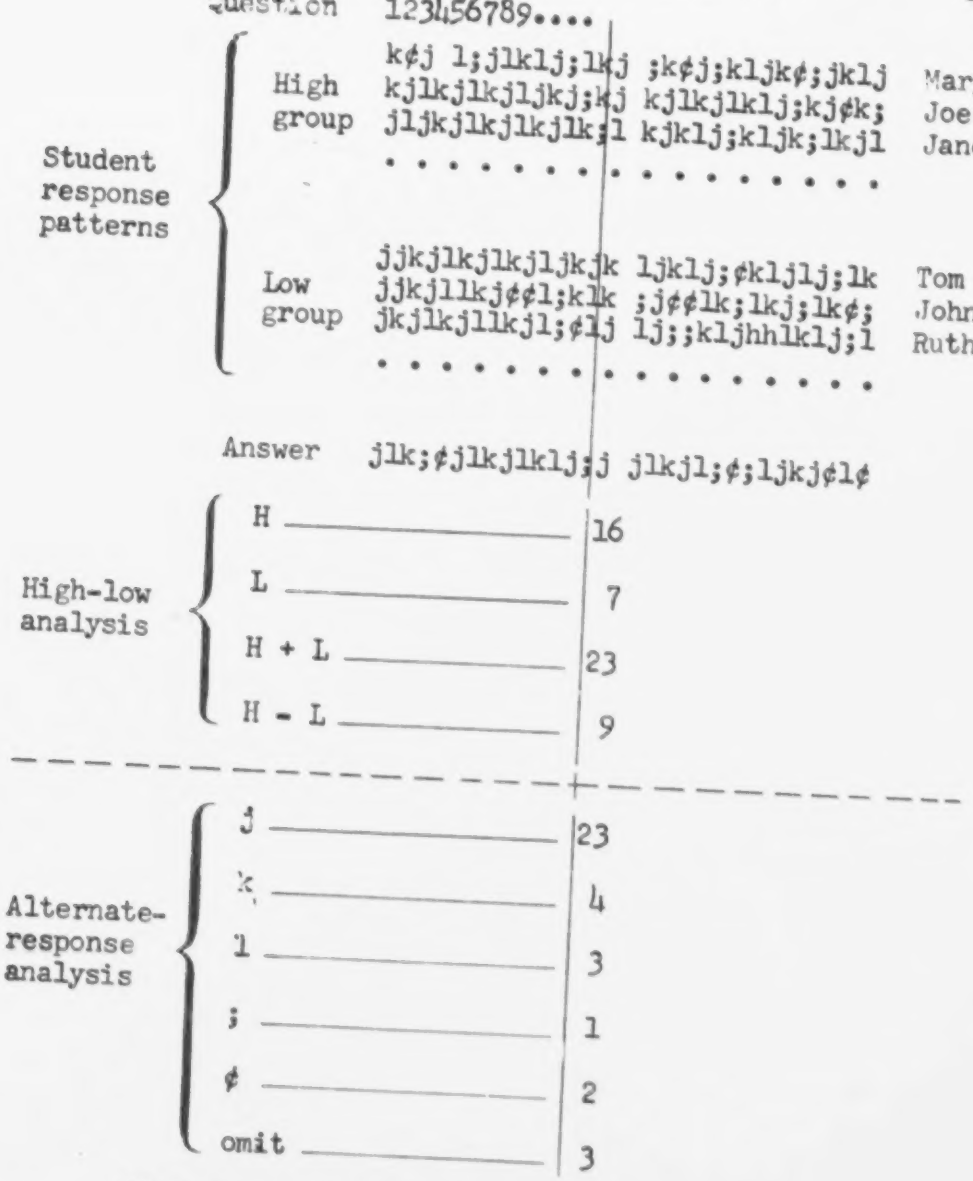


Figure 7. Sample of an Item Analysis by Typewriter With a Detailed Analysis of Item 15 [N=36]

sis by adding a row for each possible response choice below the "High-Low" Analysis. One counts and records the number of responses for each choice or omitted item. One can then discover the most common errors of the class.

This method can also be used to give error analysis information. With the vertical straightedge in position it is possible to circle or underline in red the incorrect responses. Since each row corresponds to a single student, individual help can be given as

needed to each student on the specific topics or areas covered by each item.

In other words, with a little preplanning and with *one* handling of the test papers it is possible to use this typewriter method to derive a great amount of useful information. Other kinds of interpretations will suggest themselves as one uses this technique.

VIII. Classroom Interpretation of Test Scores

IF A TEST OR TEST BATTERY is worth administering to all of the students in a class, school, school system, or State, the results should be reported to the students, teachers, administrators, and parents. The parents should be most interested in the meaning of the test scores for their students.

Before the day of the test the students should have been notified that the test was coming, told the purposes of the test, and show how the results might be interpreted. Afterwards, since they have been involved in preparation for the test and have spent a number of hours diligently marking their answer sheets, the results certainly should be explained to them as soon as possible.

The larger the testing program, the longer the interval between the administration of the tests and the report of results. However, automatic data processing is now available which provides an accurate and economical method for making results available to the schools within a maximum of 3 to 4 weeks. These reports often include list reports of scores by class, grade, building, city, or State, together with individual press-on label reports for the student's cumulative record folder, the teacher's grade book, and the student's interpretive leaflet.

As soon as the test results are available, the teachers should be given an interpretation of the scores by a qualified principal, by one of the school counselors, or by the guidance director of the school system. Staff meetings for this purpose can include presentations of meaningful interpretations of the results as related to the school, as well as suggestions for interpreting the results to the students.

The next step is to explain the results to the students. One way to accomplish this is to have the counselor or teacher explain to *each* student individually what the scores mean and how they are good measures of his strengths and weaknesses. However, such a procedure is usually not an efficient use of either the teacher's or the counselor's time.

One procedure which has been successful is to have the teacher or counselor make a general explanation to a whole class. First, it

should be pointed out to the students that the results obtained from a battery of tests is a private matter. A student is under no obligation to show his results to anyone nor should he ask to see the test results of others. Such a statement may prevent student embarrassment.

Since the students may have forgotten the types of items in the subparts of a single test or each major portion of a testing program, which may have included a scholastic aptitude test and several achievement tests, it would be appropriate to review again the purpose of each test and recall sample items of each. This may help the students to relate the type of test to their own scores. Of course, with a long standardized test battery, it would not be practical or worth while to examine each test item with the students.

The next step is to explain to the students the way in which the test scores are presented for interpretation. Results may be reported as raw scores, scaled scores, percentiles, age-equivalents, grade-equivalents, or stanines. The manual which accompanies each test explains the kinds of scores which are available and how they are derived. Each teacher should have her own copy of the interpretation manual for each test used in a testing program. Information as to how the scores are derived and their meanings will be of interest for many of the students.

The class should be told that a "national" percentile of a standardized test indicates the percentage of individuals in the group of students used to establish the test norms who made scores below that of the student. For example, a student at the 75th percentile scored better than 75 percent of the students of the norming population. It does *not* mean that the student missed only 25 percent of the test items.

It should also be pointed out that, because of errors of measurement, one should think of the reported percentiles as a *band*, rather than a particular *point* of, say, 75. That is, with a reported percentile of 75 the true score of the student might lie somewhere between 70 and 80. Further, differences of percentiles between two achievement tests should not be considered significant unless the scores are separated enough so that these bands would not tend to overlap. For example, suppose that a student ranked at the 73d percentile in mathematics and 77th percentile in English. If the interpretation manual indicates that there is an "error of measurement" of 5 percentile points at this part of the distribution of scores, then the probability is 2 out of 3 that a band of scores from 68 to 78 includes the mathematics score and a band

of scores from 72 to 82 contains the English score. Since the scores of 73 and 78 occur in *both* bands, it is quite possible that with a second administration of similar tests the percentiles could be reversed. In other words, for these tests at this part of the distribution the difference in percentiles is *not* significant.

An explanation should be given to the students of the meaning of "norms." One can explain, for example, that a sample consisting of a cross section of students of a grade and age similar to themselves was selected from all parts of the country, from industrial and agricultural areas, from large and small schools, and from prosperous suburban and crowded urban classrooms, that all of these students were given the same tests, and that a percentile or some other type of derived score was computed and these results published as the "national norms." It should be pointed out that if tests from several different publishers are used as a part of a testing battery that the "national norms" were established on different samples of students which might account for certain small unexpected differences.

"Local norms" are established on the basis of the same students within a smaller area—a State, county, city, or school. If norms are constructed for the same students for several tests at the same time, they will be comparable. Also, these local norms compare the student with his peers in his own community.

One of the most meaningful ways to analyze the results of a test battery is by means of a profile chart. This can be explained by constructing a sample profile on the blackboard, by preparing a large chart ahead of time, by using a flannel board on which the scores and lines may be placed, or by using a ruled metal board with magnetic spots to represent the scores and lines to join them.

After these explanations, each student should be given an unmarked profile sheet and a copy of his scores. Some test publishers furnish such profile sheets with their tests. If not, such sheets can be easily drawn and duplicated.

The plotting of profiles for a class would require much teacher time. Rather than distributing individually plotted profiles, it is quite acceptable, under proper supervision, to let each student plot the marks corresponding to his national percentile scores on the proper line for each test and then join the marks with a solid black line to form his own profile. This solid line represents his positions relative to the *national* norms. It is then possible for the student to see the peaks and the valleys which indicate his own relative strengths and weaknesses. Those points on the pro-

file which are low will indicate the weak areas which may need further study while the high points will emphasize the apparent strengths and may suggest several areas for future specialized study. If the student is planning post-high-school education, it may be helpful to learn how he compares nationally with those

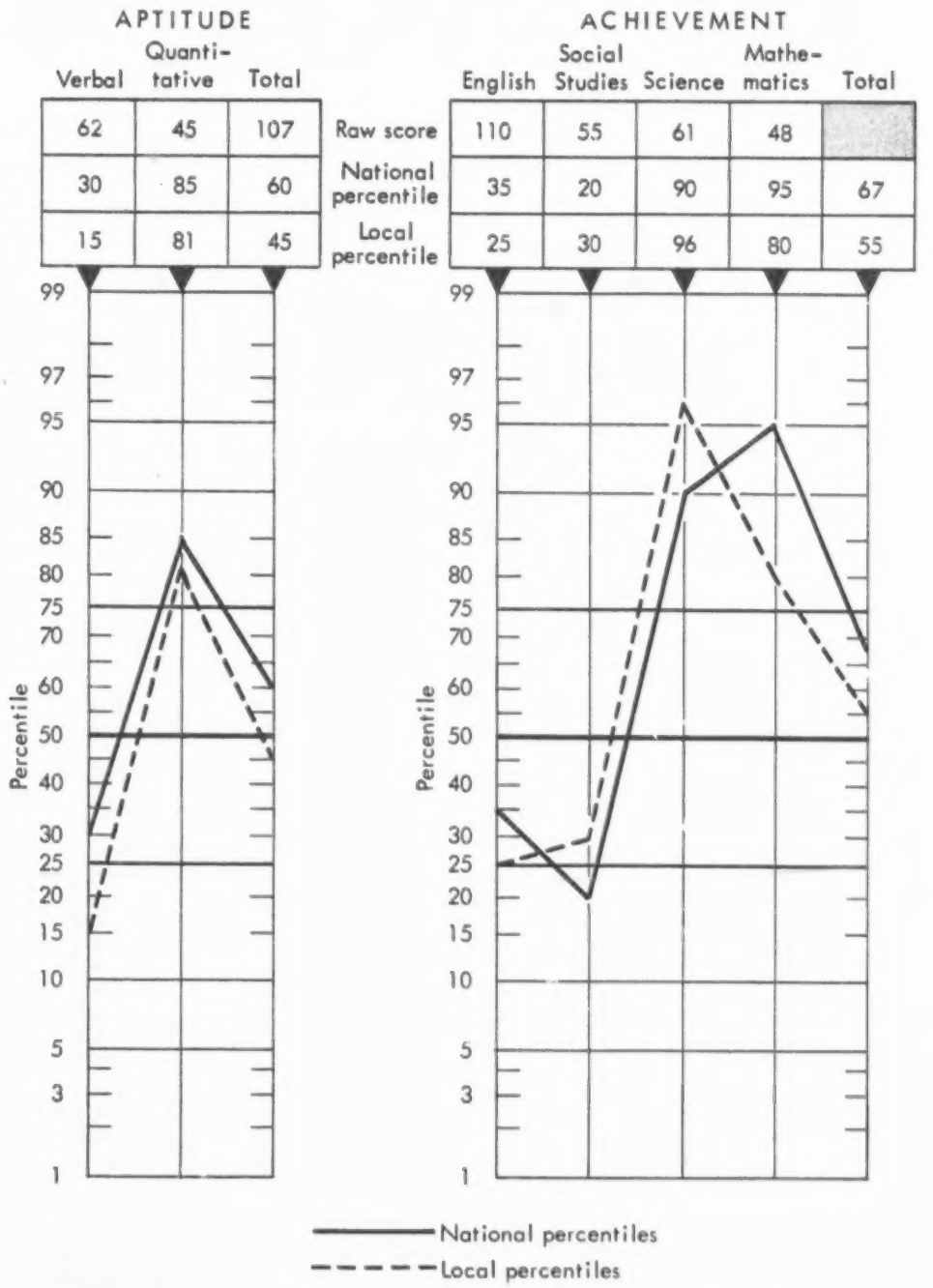


Figure 8. Profile Sheet for Sam Smith of Patunka High School, Grade 10

with whom he will be competing. Certainly, if he has such ambitions, he should be especially concerned in those areas in which he falls much below the median, or 50th percentile.

If available, each student may be given his standing in terms of *local* percentile norms. These points can be plotted and connected by a broken line or drawn in color. This line will show his relative standing as compared with his own peers or those of the surrounding community. These results may be helpful if he plans to remain in the same geographical region and compete for jobs in the local labor market. Those students included in the local norms are the types of people with whom he probably will be competing.

According to the student profile in figure 8, Sam Smith is a little better than average in overall aptitude. His total raw score of 107 places him at the the 60th percentile in terms of national norms, or in the top 40 percent. In terms of local percentiles he stands somewhere in the middle of the distribution of his classmates.

In verbal aptitude, nationally, Sam ranks in the bottom third of students of his own age and grade. On local norms he is in the lowest fifth of his class. However, in quantitative aptitude he exceeds more than four-fifths of his peers, for he scores at approximately the 85th percentile on national norms and the 81st percentile on local norms. If one allows for an error of measurement of 5 percentile points, he is still in the top quarter both nationally and locally. When one combines the verbal and quantitative aptitude scores to obtain a total score, Sam appears to be a little above the average on the national norms and a little below average on the local norms.

The strongest achievement area for Sam is *not* English or social studies. He is in about the lowest third of the class in these subjects and may need extra help. However, it is not too surprising that Sam is weak in these areas, since his verbal aptitude is low, and research shows a relationship between verbal ability and success in English and social studies.

Sam's greatest strength seems to be in the mathematics and science subjects. He is in the top 10 percent in science on both national and local norms. One may question why in mathematics he stands at the 95th percentile on the national norms and at the 80th percentile on the local norms. Test scores cannot tell us the reasons, but they can point out areas which need further thought or investigation. One explanation might be that at Patunka High School many of the students are naturally good in mathematics. Another reason might be that an unusually dedicated mathematics

teacher has motivated the students to do much better than students of preceding years. All of those who equalled or exceeded Sam's score would rank at the 95th percentile or better on the *national* norms. On local norms about one-fifth of the students are better than Sam, so his local percentile drops a few points.

The total achievement score places Sam in the top third on national norms and in the upper half on local norms. No total raw score is given in the boxhead for the achievement tests, since it would be meaningless because of the different test lengths. The total achievement percentiles were computed by methods outlined in the test manual for the tests used.

After the general class discussion by a trained teacher or by the counselor of the school, and after the students have plotted their own profiles, an opportunity should be given for any general class questions concerning the meaning of the scores. At the conclusion of the discussion, the students should be encouraged to make appointments for individual consultations concerning the test results. By explaining to the class as a whole the general meaning of these test results, many hours of individual explanations and interpretations will be saved and the students themselves will be better informed and better prepared for individual counseling.

IX. Interpretation of Test Results to Parents

Individual Conferences

How much general information or how many details concerning the test results should be given to a parent during a conference about his child? As much information and as many details should be given to the parent as the principal, the counselor, or the teacher believes can be understood and used properly. This does not mean that test results should be considered "top secret" but that there is nothing gained by making available information which may be misinterpreted.

For example, as a general policy most schools will not indicate to a child or to his parents the exact level of the child's intelligence quotient (or IQ) because the concept of the IQ is misunderstood by many people. It is even difficult to get psychologists and educators to agree upon a definition of intelligence acceptable to all. Some parents believe that an IQ is a precise measure of their child's ability, rather than an indicator of the approximate range of values in which the child's IQ lies, and will praise or condemn on the basis of this single number.

Some people talk about a verbal intelligence, a mathematical or quantitative intelligence, a mechanical or manipulative intelligence, a spatial intelligence, and so on. Studies which have been made in the past have attempted to identify certain "factors" which together seem to make up intelligence. These factors occur in varying amounts in different individuals. Studies have indicated that success in certain occupations may be reasonably expected when an individual possesses specified minimums of some of these factors. No tests have been developed to date which will measure accurately and reliably the motivation or drive of a student, and it is quite possible that students who rank below statistically derived cutoff scores on some tests, say, in mathematics or English, might do well in these same areas. However, the odds are against such an accomplishment, and the student and parent should be aware of these facts as they make decisions concerning future education and occupational preparation.

It is considered legitimate to indicate to the parent that his child seems to have unusual ability, as shown by an intelligence

test or a scholastic aptitude test, and to suggest that his area of strength is in the verbal areas rather than quantitative areas or vice versa. On the other hand, it would be acceptable to indicate that a child who is in the lower quarter of the student population in ability seems to be doing as well in his work as can be expected.

Similarly it can be pointed out to a parent that, if the child seems to be particularly gifted and the results of achievement tests seem to indicate mediocre attainment, the child is not working up to his expected capacity and he should be encouraged to use his abilities better. If the gifted child comes from a family that in the past has not made learning opportunities available, perhaps the parents can be shown the possibilities of their child and encouraged to help him gain educational experiences and materials from sources outside the school.

It is important also to point out contradictions in the data or conflicts between test results and the observations of teachers or counselors. The child's cumulative record may suggest that further individual testing may be needed. The parent should understand the reasons for additional testing.

Most parents are eager to learn the level of ability of their child. They may have pertinent suggestions to offer as to why the child either is not doing as well as expected or better than expected. Such views should be incorporated in the student's cumulative record for future reference.

Sometimes questions arise concerning a child who does not seem to be meeting even the minimum standard for his grade, although his test scores suggest that he has the ability to be at the top of his class. If the parent cannot give reasonable explanations for this, then the situation requires further investigation by the counselor or other members of the pupil personnel staff.

Group Conferences

The ideal way to explain a school's testing program and to present a student's test results to his parents is by an individual parent conference. However, with the current student-counselor ratio of 500 to 1, or greater, in many of our schools, there are not enough hours in a day to carry out such a procedure. Most counselors do not have the time to make sure that each family understands the philosophy of testing, the reasons certain tests were selected, and the meaning of the scores.

When long individual conferences are not possible, the next best thing is to have a number of the parents meet, at which time they may be given general background information. Such gatherings

might be one of the regular parent-teacher association meetings or a meeting called for this specific purpose. In some schools it has been found convenient and helpful to invite the parents of a single class for the discussion of tests and the presentation of basic information. However, in a large elementary or high school, it would take many weeks for trained personnel to reach all parents in this manner. Because of the interest in testing, it has been found that parents respond enthusiastically to the announcement of an opportunity to discuss different kinds of tests and how they may be used.

The question arises as to the best procedure for a large group discussion of tests. Several approaches are possible. In one type of program, the first half of the scheduled time may be devoted to a discussion of the current testing program by a counselor or testing specialist. A summary may be given of the complete testing program of the school or city. The purpose of each kind of test at each grade level is presented. Parents are told that tests may be administered in the fall to help the teacher learn quickly the level of ability of each of his students. Further, some tests are administered in both the fall and the spring in a few grades or classes in order to compare the effectiveness of new teaching methods. Other tests are administered at various times in some senior high schools to determine possible scholarship winners for colleges and universities.

After this general discussion, the particular tests which have just been completed by the students are described. Sample items may be shown, either by distributing a mimeographed sheet containing the illustrative sample items which were used by the students with the tests, by using slides or an overhead projector, or by using an opaque projector. It is helpful to show the parents some of the variety of forms in which the objective test items occur and to emphasize that one does *not* test for facts alone, as popular writers imply, but that one can test for basic skills and the ability to reason as well. Many of the older parents were not "subjected" to such skillfully designed standardized tests as are available today. When they were in school, citywide and statewide programs did not exist or were just being developed.

The length of a test can be discussed. Other things being equal, the longer a test, the more dependable the results. Many times a test requiring 45 minutes is preferable to another test with a similar title which requires only 10 minutes. The testing specialist says that the longer test is more "reliable," i.e., if, within a few days, a student took the same test or a parallel form, he would

receive approximately the same score. (A parallel form of a test is one which was constructed by the same author according to the same table of specifications by using items from the original pool of questions.)

One can explain the meaning of the various kinds of norms which have been used and briefly explain in nontechnical terms how they were derived and what they mean. Sample profiles can be distributed or displayed on a screen so that all present can follow their interpretation. A profile similar to the one described in the preceding section would be helpful.

After this presentation, it is possible to open the meeting to questions. The chairman may accept questions from the floor. Another method is to distribute cards early in the program so that those present may write their questions. As soon as these cards are collected, it is possible for the moderator to group related or duplicating questions quickly and to determine the difficulties or lack of understandings among those present. Specific questions can be read and answered without embarrassment to any parent.

Another way to prepare for a parent meeting is to circulate to the parents via the students a series of possible questions and topics concerning testing and the meaning of the results. The parents should be asked to check those questions which are of most concern or of most interest to them and to return the forms at least 2 weeks before the scheduled meeting. A quick tally will indicate those topics which should be included in the program. Such a questionnaire also helps the parent to think about the discussion topics and to formulate questions for the discussion period.

In a third type of meeting, during which anonymous case studies may be presented, the first part of the allotted time is devoted to a short discussion of the tests used. Then a distribution is made to all parents of a letter-sized paper folded over once and stapled. All persons are cautioned *not* to remove the staple until told to do so.

Without opening the sheet, each person can read on the visible portion of the paper all of the information available concerning one anonymous student. Information about the student would include his attendance record, course grades, extracurricular activities, interest measures, test scores, and general family and community background.

Each person present would be asked to think about the rela-

tionship of the test scores to all other information. These questions could be raised:

1. What would each parent tell this student on the basis of the information available?
2. Is all of the information necessary?
3. Would the test scores alone be sufficient to indicate student needs?
4. Is the information adequate to assist the student without the test scores?
5. What is meant by national and local norms? (These could be explained.)

A few volunteers from the audience might be willing to attempt an interpretation of the results and suggest proper action. After everyone has exhausted his ideas, the sheets may be opened so that everyone may see one of several possible professional interpretations based upon the listed facts.

Programs of this case study type have proved worth while as a basis for group discussion. The inclusion of several contrasting cases can be helpful. The use of a different colored paper for each case will assure that all are talking about the same one.

Timing is an important consideration in discussions of tests with parents. Such meetings are most helpful to the parents if they occur before or at the same time as the distribution of test results to the students. Advance publicity through the local press announcing the early availability of test results and explanatory meetings can also prove helpful.

There is some difference of opinion as to whether test results should be sent home to the parents. Certainly, no test results should be distributed by mail or taken home by the child unless accompanied by a short description of the test and a simple explanation of the meaning of the results. This principle is frequently ignored. Many times a child has brought home a piece of paper with numbers, but with no indication as to whether they represent low scores or high scores, bad scores or good scores.

The story is told that when one parent was given a test report by his child he asked, "What's this?" The child replied, "Tests!" This was the total amount of explanation available to the parent!

A cartoon [1] which appeared in 1955 illustrates one possible misinterpretation of test results. A distraught mother was shown calling the doctor about her son who had just brought home a

[1] By Gardner Rea in *LOOK* magazine, copyright 1955, by Cowles Magazines, Inc. Also in *Test Service Bulletin*, No. 54, December 1959, "On Telling Parents About Test Results," by James H. Ricks, Jr. New York: The Psychological Corporation, p. 3.

card indicating that he had an IQ of 105. She wanted to know whether he should be put right to bed. Obviously, no information had been sent to explain the meaning of this number to the parent. On the other hand, some schools send home with the child a 4-page brochure explaining the purposes of the tests, the meaning of the scores, and inviting the parent to make an appointment for a conference if further information is desired.

Often the schools fail to take advantage of the local press as a means of informing the parents concerning a schoolwide test either impending of completed. A Florida county a few years ago gave much publicity to the importance of a certain testing program for all students. As a result, the attendance at school that day was the best of the year. No parent wanted his child to miss out on tests which could help him.

The press is willing to inform the public concerning the activities of the school, including information about tests. If given the material and the proper assistance, newspapers will print a worthwhile discussion, including a description of the tests and the implications of test results. Such publicity can arouse the interest of the parents and encourage them to come to a parent-teacher meeting to learn more about tests.

X. Selected References

ADAMS, GEORGIA S. *and* TORGERSON, THEODORE L. *Measurement and Evaluation for the Secondary School Teacher with Implications for Corrective Procedures.* New York: The Dryden Press, 1956. 658 p.

AHMANN, J. STANLEY *and* GLOCK, MARVIN D. *Evaluating Pupil Growth.* Boston: Allyn and Bacon, 1958. 605 p.

——— ——— *and* WARDEBERG, HELEN L. *Evaluating Elementary School Pupils.* Boston: Allyn and Bacon, Inc. 1960. 435 p.

AMERICAN COUNCIL ON EDUCATION, COMMITTEE ON MEASUREMENT AND EVALUATION. *College Testing, A Guide to Practices and Programs.* Washington, D.C.: American Council on Education, 1959. 190 p.

ANASTASI, ANNE. *Psychological Testing.* 2d ed. New York: The Macmillan Co., 1961. 657 p.

ANDERSON, SCARVIA B.; KATZ, MARTIN; *and* SHIMBERG, BENJAMIN. *Meeting the Test.* New York: Scholastic Book Services, 1963. 184 p.

BAUERNFEIND, ROBERT H. *Building a School Testing Program.* Boston: Houghton Mifflin Co., 1963. 343 p.

BEAN, KENNETH L. *Construction of Educational and Personnel Tests.* New York: McGraw-Hill Book Co., Inc. 1953. 231 p.

BERDIE, RALPH F.; LAYTON, WILBUR L.; SWANSON, EDWARD O.; *and* HAGENAH, THEDA. *Counseling and the Use of Tests, A Manual for the State-Wide Testing Programs of Minnesota.* Minneapolis, Minn.: University of Minnesota, 1959. 178 p.

——— ——— ——— ——— *Testing in Guidance and Counseling.* New York: McGraw-Hill Book Co., Inc., 1963. 288 p.

BRADFIELD, JAMES M. *and* MOREDOCK, H. STEWART. *Measurement and Evaluation in Education.* New York: The Macmillan Co., 1957. 509 p.

CHAUNCEY, HENRY *and* DOBBIN, JOHN E. *Testing: Its Place in Education Today.* 1st ed. New York: Harper & Row, Publishers, Inc., 1963. 224 p.

COLLEGE ENTRANCE EXAMINATION BOARD. *Manual of Freshman Class Profiles.* Box 592, Princeton, N.J.: College Entrance Examination Board.

CRONBACH, LEE J. *Essentials of Psychological Testing.* 2d ed. New York: Harper & Bros., 1960. 650 p.

DAILEY, JOHN T. *and* SHAYCOFT, MARION F. *Types of Tests in Project Talent.* Washington: U.S. Government Printing Office, 1961. 62 p. (U.S. Office of Education, Cooperative Research Monograph, No. 9 OE–25014.)

DAVIS, FREDERICK B. *Educational Measurements and Their Interpretation.* Belmont, Calif.: Wadsworth Publishing Co., 1964.

——— Item Selection Techniques. *Educational Measurement.* Washington: American Council on Education, 1951. p. 266–328.

DIEDERICH, PAUL B. *Short-cut Statistics for Teacher-made Tests.* Princeton, N.J.: Educational Testing Service, 1960. 44 p. (Evaluation and Advisory Service Series, No. 5)

DOPPELT, JEROME E. How Accurate Is a Test Score? *Test Service Bulletin,* No. 50, New York: The Psychological Corporation, June 1956. p. 1–3.

——— *and* SEASHORE, HAROLD G. How Effective Are Your Tests? *Test Service Bulletin,* No. 37. New York: The Psychological Corporation, June 1949. p. 4–10.

DUROST, WALTER N. The Characteristics, Use, and Computation of Stanines. *Test Service Notebook,* No. 23. New York: Harcourt, Brace & World, Inc., 1961. 6 p.

——— *How To Tell Parents About Standardized Test Results. Test Service Notebook,* No. 26. Tarrytown, N.Y.: Harcourt, Brace & World, Inc., 1961. 4 p.

——— Why Do We Test Your Children? *Test Service Notebook,* No. 17, New York: Harcourt, Brace & World, Inc., 1956. 4 p.

——— *Manual for Interpreting Metropolitan Achievement Tests.* Primary I Through Advanced. New York: Harcourt, Brace & World. 1963.

——— *and* PRESCOTT, GEORGE A. *Essentials of Measurement for Teachers.* New York: Harcourt, Brace & World, Inc., 1962. 167 p.

FINDLEY, WARREN G., ed. *The Impact and Improvement of School Testing Programs.* 62d Yearbook, Part II. 5835 Kimbark Avenue, Chicago: National Society for the Study of Education, 1963. 304 p.

FREEMAN, FRANK SAMUEL. *Theory and Practice of Psychological Testing.* 3d ed. New York: Holt, Rinehart and Winston, Inc., 1962. 697 p.

FROEHLICH, CLIFFORD P. *and* HOYT, KENNETH B. *Guidance Testing and Other Student Appraisal Procedures for Teachers and Counselors.* 3d ed. Chicago: Science Research Associates, Inc., 1959. 438 p.

FURST, EDWARD J. *Constructing Evaluation Instruments.* New York: Longmans, Green and Co., 1958. 334 p.

GANNON, F. B. *and* TELSCHOW, EARL. *Tests and Interpretations*—a teacher's handbook. Rochester, N.Y.: City School District, 1960. 31 p.

A Glossary of Measurement Terms, Ninety-six Concepts Which Constitute a Basic Vocabulary in Evaluation and Testing. Del Monte Research Park, Monterey, Calif.: California Test Bureau, 1959. 16 p.

GOLDMAN, LEO. *Using Tests in Counseling.* New York: Appleton-Century-Crofts, Inc., 1961. 434 p.

GOODENOUGH, FLORENCE L. *Mental Testing: Its History, Principles, and Applications.* New York: Rinehart and Co., Inc., 1949. 609 p.

GOSLIN, DAVID A. *The Search for Ability; Standardized Testing in Social Perspective.* Volume 1 of a Series on the Social Consequences of Ability Testing. New York: Russell Sage Foundation, 1963. 204 p.

GREEN, JOHN A. *Teacher-Made Tests.* New York: Harper & Row, Publishers, 1963. 141 p.

GREENE, EDWARD B. *Measurements of Human Behavior.* Rev. ed. New York: The Odyssey Press, 1952. 790 p.

HART, IRENE. Using Stanines To Obtain Composite Scores Based on Test Data and Teachers' Ranks. *Test Service Bulletin,* No. 86. Tarrytown, N.Y.: Harcourt, Brace & World, Inc., 1957. 4 p.

HUMPHREYS, J. ANTHONY; TRAXLER, ARTHUR E.; *and* NORTH, ROBERT D. *Guidance Services.* 2d ed. Chicago: Science Research Associates, Inc., 1960. 414 p.

JACOBS, JAMES N. Aptitude and Achievement Measures in Predicting High School Academic Success. *The Personnel and Guidance Journal,* 3:334–341, January 1959. *Also reprinted* as *Test Service Bulletin,* No. 94. Tarrytown, N.Y.: Harcourt, Brace & World, Inc. 6 p.

JORDAN, A. M. *Measurement in Education, An Introduction.* New York: McGraw-Hill Book Co., Inc., 1953. 533 p.

KATZ, MARTIN R. *Selecting an Achievement Test; Principles and Procedures.* (Evaluation and Advisory Service Series, No. 3) Princeton, N.J.: Educational Testing Service, 1958. 32 p.

KENT AREA GUIDANCE COUNCIL. *A Proposed Twelve-Year Testing Program.* Columbus, Ohio: Ohio Scholarship Tests, State Department of Education, March 1959. 57 p.

LENNON, ROGER T. Testing in the Secondary School. *Test Service Notebook, No. 20.* Tarrytown, N.Y.: Harcourt, Brace & World, Inc., 1957. 4 p.

———, *et al.* A Glossary of 100 Measurement Terms. *Test Service Notebook,* No. 13. Tarrytown, N.Y.: Harcourt, Brace & World, Inc. p. 6.

LINDQUIST, E. F., *ed. Educational Measurement.* Washington, D.C.: American Council on Education, 1951. 820 p.

LINDVALL, C. M. *Testing and Evaluation: An Introduction.* New York: Harcourt, Brace & World, Inc., 1961. 264 p.

LYMAN, HOWARD B. *Test Scores and What They Mean.* Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1963. 223 p.

MCCABE, GEORGE E. Test Interpretation in the High School Guidance Program. *Test Service Bulletin, No. 93.* Tarrytown, N.Y.: Harcourt, Brace & World, Inc. p. 1–3.

MCLAUGHLIN, KENNETH F. How Is a Test Built? *Understanding Testing.* Washington: U.S. Government Printing Office, 1962. p. 4–7. (U.S. Office of Education, OE–25003.) *Also reprinted* as a *Test Service Notebook, No. 25.* Tarrytown, N.Y.: Harcourt, Brace & World, Inc.

———, *ed. Understanding Testing.* Washington: U.S. Government Printing Office, 1962. 24 p. (U.S. Office of Education, OE–25003.)

NOLL, VICTOR H. *Introduction to Educational Measurement.* Boston: Houghton Mifflin Co., 1957. 437 p.

REMMERS, H. H. *and* GAGE, N. L. *Educational Measurement and Evaluation.* Rev. ed. New York: Harper & Bros., 1955, 650 p.

——— ——— *and* RUMMEL, J. FRANCIS. *A Practical Introduction to Measurement and Evaluation.* New York: Harper & Bros., 1960. 370 p.

RICKS, JAMES H., JR. On Telling Parents About Test Results. *Test Service Bulletin, No. 54,* December 1959. New York: The Psychological Corporation. 4 p.

ROSS, C. C. *and* STANLEY, JULIAN C. *Measurement in Today's Schools.* 3d ed. New York: Prentice-Hall, Inc., 1954. 485 p.

ROTHNEY, JOHN W. M.; DANIELSON, PAUL J.; *and* HEIMANN, ROBERT A. *Measurement for Guidance.* New York: Harper & Bros., 1959. 378 p.

RUSSELL, ROGER W. *and* CRONBACH, LEE J. Report of Testimony at a Congressional Hearing (to the Senate Committee on Labor and Public Welfare on Feb. 27, 1958). *The American Psychologist* 13:219–220, March 1958.

SCHWARTZ, ALFRED *and* TIEDEMAN, STUART C. *Evaluating Student Progress in the Secondary School.* 1st ed. New York: Longmans, Green and Co., 1957. 434 p.

SEASHORE, HAROLD G. Methods of Expressing Test Scores. *Test Service Bulletin, No. 48.* New York: The Psychological Corporation, January 1955. p. 7–9.

SEGEL, DAVID; WELLMAN, FRANK E.; *and* HAMILTON, ALLEN T. *An Approach to Individual Analysis in Educational and Vocational Guidance.* Washington: U.S. Government Printing Office, 1958. 39 p. (U.S. Office of Education, Bulletin 1959, No. 1.)

STODOLA, QUENTIN. How One School System Records and Interprets Test Scores: A Do-It-Yourself Kit for Teachers. *Test Service Bulletin, No. 89.* Tarrytown, N.Y.: Harcourt, Brace & World, Inc., 1958. 6 p.

———. *Making the Classroom Test; A Guide for Teachers.* Princeton, N.J.: Educational Testing Service, 1959. 28 p. (Evaluation and Advisory Service Series, No. 4)

Testing Guide for Teachers. Prepared by the Technical Subcommittee of the Independent Schools Advisory Committee. New York: Educational Records Bureau, 1961. 43 p.

THOMAS, R. MURRAY. *Judging Student Progress.* New York: Longmans, Green and Co., 1954. 421 p.

THORNDIKE, ROBERT L. *and* HAGEN, ELIZABETH. *Measurement and Evaluation in Psychology and Education.* 2d ed. New York: John Wiley & Sons, Inc., 1961. 602 p.

TRAVERS, ROBERT M. W. *Educational Measurement.* New York: The Macmillan Co., 1955. 420 p.

TRAXLER, ARTHUR E. *Techniques of Guidance.* Rev. ed. New York: Harper & Bros., 1957. 374 p.

——— JACOBS, ROBERT; SELOVER, MARGARET; *and* TOWNSEND, AGATHA. *Introduction to Testing and the Use of Test Results in Public Schools.* New York: Harper & Bros., 1953. 113 p.

TRIGGS, FRANCES ORALIND. *Reading: Its Creative Teaching and Testing: Kindergarten Through College.* Mountain Home, N. C.: Frances Oralind Triggs, Chairman, The Committee on Diagnostic Reading Tests, Inc., 1960. 150 p.

TYLER, LEONA E. *Tests and Measurements.* Foundations of Modern Psychology Series. Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1963. 116 p.

WANDT, EDWIN *and* BROWN, GERALD W. *Essentials of Educational Evaluation.* New York: Henry Holt and Co., Inc., 1957. 117 p.

WESMAN, ALEXANDER G. Aptitude, Intelligence, and Achievement. *Test Service Bulletin, No. 51,* December 1956. New York: The Psychological Corporation. p. 4–6.

———. Expectancy Tables—A Way of Interpreting Test Validity. *Test Service Bulletin, No. 38.* New York: The Psychological Corporation, December 1949. p. 11–15.

WILLEY, CLARENCE F. Simplified Item Analysis. *Public Personnel Review* 14:24–25, January 1953.

WOMER, FRANK B. Initiating a Testing Program. *The Elementary School Journal* 57:193–97, January 1957. *Also reprinted* as *A Test Service Bulletin, No. 14.* Boston: Houghton Mifflin Co., 3 p.

WRIGHTSTONE, J. WAYNE; JUSTMAN, JOSEPH; *and* ROBBINS, IRVING. *Evaluation in Modern Education.* New York: American Book Co., 1956. 481 p.

☆ U.S. GOVERNMENT PRINTING OFFICE : 1965—O-791-247